

# Approximations of the allelic frequency spectrum in general supercritical branching populations

BENOIT HENRY<sup>1,2</sup>

## Abstract

We consider a general branching population where the lifetimes of individuals are i.i.d. with arbitrary distribution and where each individual gives birth to new individuals at Poisson times independently from each other. In addition, we suppose that individuals experience mutations at Poissonian rate  $\theta$  under the infinitely many alleles assumption assuming that types are transmitted from parents to offspring. This mechanism leads to a partition of the population by type, called the allelic partition. The main object of this work is the frequency spectrum  $A(k, t)$  which counts the number of families of size  $k$  in the population at time  $t$ . The process  $(A(k, t), t \in \mathbb{R}_+)$  is an example of non-Markovian branching process belonging to the class of general branching processes counted by random characteristics. In this work, we propose methods of approximation to replace the frequency spectrum by simpler quantities. Our main goal is study the asymptotic error made during these approximations through central limit theorems. In a last section, we perform several numerical analysis using this model, in particular to analyze the behavior of one of these approximations with respect to Sabeti's Extended Haplotype Homozygosity [18].

*MSC 2000 subject classifications:* Primary 60J80; secondary 92D10, 60J85, 60G51, 60K15, 60F05.

*Key words and phrases.* branching process – splitting tree – Crump–Mode–Jagers process – linear birth–death process – Central Limit Theorem.

## 1 Introduction

In this paper, we consider a general branching population where the lifetimes of the individuals and their reproductions processes are i.i.d. Moreover, we assume that their lifetimes are distributed according to an arbitrary probability distribution  $\mathbb{P}_V$  and that the births occur, during their lifetime, according to a Poisson process with rate  $b$ . The tree underlying this dynamics is called a splitting tree. This class of random trees was introduced in [11] by Geiger and Kersting and has been widely studied in the last decade [14, 15, 16].

---

<sup>1</sup>Madynes team, INRIA Nancy – Grand Est, IECL – UMR 7503, Nancy-Université, Campus scientifique, B.P. 70239, 54506 Vandœuvre-lès-Nancy Cedex, France

<sup>2</sup>LORIA – UMR 7503, Nancy-Université, Campus scientifique, B.P. 70239, 54506 Vandœuvre-lès-Nancy Cedex, France, E-mail: [benoit.henry@univ-lorraine.fr](mailto:benoit.henry@univ-lorraine.fr)

We suppose, in addition, that mutations occur on individuals and that each new mutation confers to its holder a brand new type (i.e. never seen in the population): this is the *infinitely many alleles* assumption. This allows modeling the occurrence of a new type in a population (such as a new species or a new phenotype in a given species). We also suppose that every individual inherits the type of its parent. This model leads to a partition of the population by types. The frequency spectrum of the population alive at time  $t$  is defined as the sequence of number  $(A(k, t))_{k \geq 1}$  where, for each  $k$ ,  $A(k, t)$  is the number of families of size  $k$  in the population. The famous example of Ewens sampling formula gives explicit expression for the law of the frequency spectrum [9] when the genealogy is given by the Kingman's coalescent. Other works studied similar quantities in the case of Galton-Watson branching processes (see [4] or [12]). In our model, the frequency spectrum has also been widely studied in the past [6, 7, 8, 5].

Another object of interest is the process  $(N_t, t \in \mathbb{R}_+)$  which counts the number of living individuals in the population at a given time  $t$ . This process is known as binary homogeneous Crump-Mode-Jagers process. One of the main result of the theory of such process is the law of large number which gives in our particular case that  $e^{-\alpha t} N_t$  converges almost surely to a random variable  $\mathcal{E}$  which is exponential conditionally on non-extinction (for some positive constant  $\alpha$ ).

As for  $e^{-\alpha t} N_t$ , it is also known that the quantities  $e^{-\alpha t} A(k, t)$  converge almost surely to  $c_k \mathcal{E}$ , where  $c_k$  is an explicit constant. This result can be easily obtained by conjunction of the works of [6] and [17] using the theory of general branching processes counted by random characteristics (a complete statement can be found in [8]). An alternative proof avoiding the use of the general branching processes theory can be found in [5].

It appears that the frequency spectrum  $(A(k, t))_{k \geq 1}$  is a quantity which is hard to manipulate from the probabilistic point of view (see [6, 7, 5]). This implies that such a model is inconvenient for practical applications. In this work we propose to use the laws of large numbers in order to replace  $(A(k, t))_{k \geq 1}$  by more manipulable quantities and propose to investigate the error made during this approximation. The first possible approximation is the following.

**Approximation 1:**

$$(A(k, t))_{k \geq 1} \approx (c_k)_{k \geq 1} e^{\alpha t} \mathcal{E}.$$

However, this is unsatisfactory for practical applications since the random variable  $\mathcal{E}$  is not observable at finite times. Another idea is to exploit the fact that the random variable appearing in the law of large numbers for  $A(k, t)$  and for  $N_t$  is the same. This leads to the second approximation.

**Approximation 2:**

$$A(k, t) \approx (c_k)_{k \geq 1} N_t.$$

In order to investigate the errors made during this approximation (at least asymptotically), one would like to have central limit theorems associated to the law of large numbers for the frequency spectrum. In a previous work [13], we have showed that the error in the convergence of  $e^{-\alpha t} N_t$  is of order  $e^{\alpha t/2}$  and obtained a central limit theorem for this error. An important aspect of the method introduced in [13] is that it can be used to derive CLTs for other branching processes counted by random characteristics. In particular, the main goal of this work is to obtain central limit theorems for the convergence of the frequency spectrum. We also study the Markovian cases (when  $\mathbb{P}_V$  is exponential) where we can obtain more information on the limit distribution.

The original motivation of this study (and of other works on this model [6, 7, 5]) comes from the works of Sabeti and al. [18] where the frequency spectrum is used to detect positive selection of an allele in an increasing population. More specifically, suppose that you want to detect the positive selection of an allele on a given gene. The main idea is that, under neutral evolution, the allele under consideration needs a long time to reach a high frequency in the population. Hence, if the frequency of the allele w.r.t. its age is significantly higher than the expected frequency (w.r.t. its age and under neutral growth), this anomaly would suggest a positive selection of this allele. The main problem is now to be able to estimate how old the allele is. Sabeti and al. remarked that the allelic partition can be used as a clock to estimate the age of an allele. More precisely, their study begins by selecting a small region of chromosome which characterized the presence of the allele under consideration. Now, the type of an individual, at a distance  $x$  (measured in kb) from the core region, is the sequence of  $x$  bases following the core region (excluded). As a consequence, the allelic partition of the subpopulation carrying the allele becomes thinner as  $x$  increases (because the higher  $x$  is, the higher is the probability that a mutation occurred on the sequence of  $x$  bases). Finally, the speed of fragmentation of the allelic partition, when  $x$  increases, gives clues on the age of the allele. One of the purposes of this model is to understand how the frequency spectrum evolves under neutral evolution. In this work, we discuss some aspects of this method and give some directions in order to construct rigorous tests for the positive selection (see Section 7).

The paper is organized as follows. Section 2 is devoted to the mathematical description of the model and to preliminary results which are used in the sequel. Section 3 gives the main theoretical results of this work and, in particular, a central limit theorem which allows to study the error in our proposed approximations. Section 4, 5, 6 are devoted to the proofs of Theorem 3.1, 3.3 and 3.5 respectively. Finally, in Section 7 we perform some numerical studies on the model to stress the quality of our approximation. The discussions about the method of Sabeti and al. are given in this last section. An appendix contains some technical proofs and a section which is a reminder of renewal theory.

## 2 Model and preliminaries

In this work, we consider a branching population with the following dynamic: starting with a single individual (called the *ancestor*) whose lifetime is distributed according to an arbitrary probability distribution  $\mathbb{P}_V$ , this *ancestor* gives birth to new individuals at a Poissonian rate  $b$ . Each birth event gives a single new individual. From this point, each child of the ancestor lives and gives birth according to the same mechanism independently from the other individuals in the population. This formal description can be made rigorous through the definition of a probability distribution on the set of chronological trees. For the details of such construction, we refer the reader to [14]. The first quantity of interest when studying such population is the number  $N_t$  of alive individuals in the population at a fixed time  $t$  (assuming that the time  $t = 0$  is birth-date of the ancestor). The process  $(N_t, t \in \mathbb{R}_+)$  is known as binary homogeneous Crump-Mode-Jagers process and is a simple example of non-Markovian branching process. In the sequel, we denote by  $W(t)$  the expectation of

$N_t$  conditionally on the non-extinction at time  $t$ . That is

$$W(t) := \mathbb{E}[N_t \mid N_t > 0].$$

In [14], the author shows that the random variable  $N_t$  is geometrically distributed under  $\mathbb{P}_t$  with parameter  $\frac{1}{W(t)}$ . In addition, the author of [14] showed that the Laplace transform of  $W$  can be linked to the Laplace transform of  $\mathbb{P}_V$  through the relation

$$\int_{[0, \infty)} W(s) e^{-\lambda s} ds = \frac{1}{\psi(\lambda)}, \quad \forall \lambda > \alpha,$$

where

$$\psi(x) = x - \int_{(0, \infty]} (1 - e^{-rx}) b\mathbb{P}_V(dr), \quad x \in \mathbb{R}_+, \quad (2.1)$$

and  $\alpha$  is the largest root of  $\psi$ . In particular, the Laplace transform of  $\mathbb{P}_V$  can be expressed in terms of  $\psi$ ,

$$\int_{\mathbb{R}_+} e^{-\lambda v} \mathbb{P}_V(dv) = 1 + \frac{\psi(\lambda) - \lambda}{b}. \quad (2.2)$$

In this work, we assume that  $\alpha$  is a strictly positive real number. This case is called the supercritical case and is equivalent to  $b\mathbb{E}[V] > 1$ . In the supercritical case, the real number  $\alpha$  is called the Malthusian parameter of the population because it corresponds to the mean exponential growth rate of the population. Before going further, let us remark that equation (2.2) leads easily to the following identity:

$$\int_{\mathbb{R}_+} e^{-\alpha v} \mathbb{P}_V(dv) = 1 - \frac{\alpha}{b}. \quad (2.3)$$

Many previous works demonstrate [6, 7, 8] that some properties of the splitting tree were easier to study on the tree describing only the genealogical relation between the lineages of the individuals alive at time  $t$ . For instance, in the model with mutations, the difference between two individuals in term of type lies only on the time past since their lineages has diverged. Hence, this particular genealogical tree, known as *coalescent point processes* (CPP), contains the essential information to study the allelic partition. In order to derive the law of that genealogical tree, one needs to characterize the joint law of the *times of coalescence* between pairs of individuals in the population, which are the times since their lineages have split.

In [14], the author defines an order on the set of individuals alive at a fixed time  $t$  and consider the sequence of times of coalescences  $(H_i)_{0 \leq i \leq N_t - 1}$  between two consecutive individuals (that is  $H_i$  is the time passed since the lineage of individuals  $i$  and  $i + 1$  have diverged) with the convention that the older lineage is the first one (i.e.  $H_0 = t$ ). Moreover, in [14], the author shows that the random vector  $(H_i)_{0 \leq i \leq N_t - 1}$  can be produced from a sequence  $(H_i)_{i \geq 1}$  of i.i.d. random variable stopped at its first value greater than  $t$  and such that

$$\mathbb{P}(H_1 > s) = \frac{1}{W(s)}, \quad s \in \mathbb{R}_+.$$

To summarize, given the population is still alive at time  $t$ , one can forget about the details of the splitting tree and code the genealogy by a new object called the *coalescent point process* (CPP). Its

law is the law of a sequence  $(H_i)_{0 \leq i \leq N_t-1}$ , where the family  $(H_i)_{i \geq 1}$  is i.i.d. with the same law as  $H$ , stopped before its first value  $H_{N_t}$  greater than  $t$ , and  $H_0$  is deterministic equal to  $t$  (see Figure 1).

Although we do not use directly the CPP in this work, this object allowed us to obtain [5] formulas for the moments of the frequency spectrum which are widely used in the sequel.

**Remark 2.1.** *Let  $N$  be a integer valued random variable. In the sequel we said that a random vector with random size  $(X_i)_{1 \leq i \leq N}$  form an i.i.d. family of random variables independent of  $N$ , if and only if*

$$(X_1, \dots, X_N) \stackrel{d}{=} (\tilde{X}_1, \dots, \tilde{X}_N),$$

where  $(\tilde{X}_i)_{i \geq 1}$  is a sequence of i.i.d. random variables distributed as  $X_1$  independent of  $N$ .

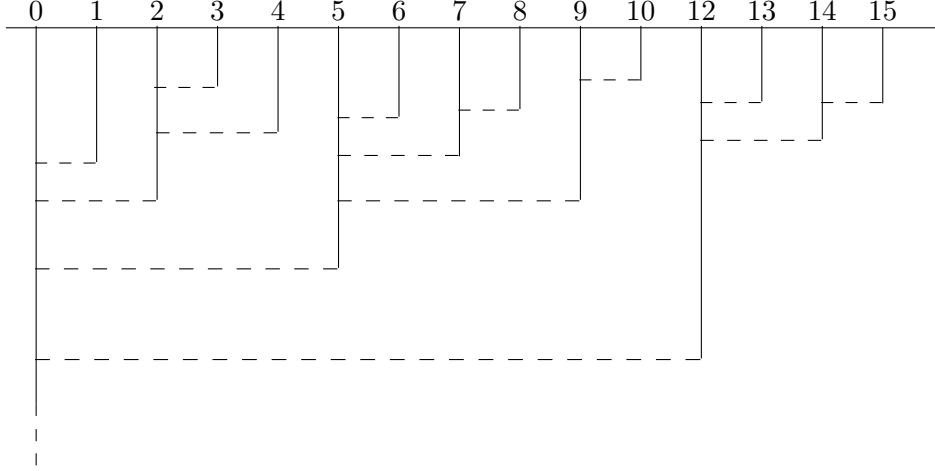


Figure 1: A coalescent point process for 16 individuals, hence 15 branches. (Image by A. Lambert)

Before going further, let us point out that if we define  $N_t$  as the first value of the sequence  $(H_i)_{i \geq 1}$  greater than  $t$ , i.e.

$$N_t = \inf\{i \geq 1 \mid H_i > t\},$$

then  $N_t$  is indeed geometric with the expected parameter. More precisely, for a positive integer  $k$ ,

$$\mathbb{P}(N_t = k \mid N_t > 0) = \frac{1}{W(t)} \left(1 - \frac{1}{W(t)}\right)^{k-1}. \quad (2.4)$$

In particular,

$$\mathbb{E}[N_t \mid N_t > 0] = W(t). \quad (2.5)$$

Moreover, it can be showed (see [17]), that

$$\mathbb{E}N_t = W(t) - W \star \mathbb{P}_V(t), \quad (2.6)$$

and

$$\mathbb{P}(N_t > 0) = 1 - \frac{W \star \mathbb{P}_V(t)}{W(t)}, \quad (2.7)$$

where

$$W \star \mathbb{P}_V(t) := \int_{[0,t]} W(t-s) \mathbb{P}_V(ds).$$

Now, let us introduce the mathematical formalism for the mutation process used in this work (this formalism comes from [5]). Since only the mutations occurring on the lineages of living individuals at time  $t$  can be observed, it follows from standard properties on Poisson point processes, that the mutation process can be defined directly on the CPP. So, let  $\mathcal{P}$  be a Poisson random measure on  $[0, t] \times \mathbb{N}$  with intensity measure  $\theta \lambda \otimes C$ , where  $C$  is the counting measure on  $\mathbb{N}$ , then the mutation random measure  $\mathcal{N}$  on the CPP is defined by

$$\mathcal{N}(da, di) = \mathbb{1}_{H_i > t-a} \mathbb{1}_{i < N_t} \mathcal{P}(di, da),$$

where an atom at  $(a, i)$  means that the  $i$ th branch experiences a mutation at time  $t-a$ . We suppose that each individual inherits the type of its parent. This rule yields a partition of the population by types. The distribution of the sizes of the families in the population is called the frequency spectrum and is defined as the sequence  $(A(k, t))_{k \geq 1}$  where  $A(k, t)$  is the number of types carried by exactly  $k$  individuals in the alive population at time  $t$ , excluding the family holding the ancestral type of the population (i.e. individuals holding the same type as the root at time 0). This last family is called *clonal*, as the ancestral type.

In the study of the frequency spectrum, an important role is played by the law of the clonal family. We denote by  $Z_0(t)$  the size of this family at time  $t$ .

To study this family, it is easier to consider the clonal splitting tree constructed from the original splitting tree by cutting every branches beyond mutations. This clonal splitting tree is a standard splitting tree without mutations, where individuals are killed as soon as they die or experience a mutation. The new lifespan law is therefore the minimum between an exponential random variable of parameter  $\theta$  and an independent copy of  $V$ . It is straightforward by simple manipulations of Laplace transforms that the Laplace exponent of the corresponding contour process is

$$\psi_\theta(x) = x - \int_{(0,\infty]} (1 - e^{-rx}) \Lambda_\theta(dr) = \frac{x\psi(x + \theta)}{x + \theta}.$$

We denote by  $W_\theta$  the corresponding scale function. This leads to,

$$\mathbb{P}(Z_0(t) = k \mid Z_0(t) > 0) = \frac{1}{W_\theta(t)} \left(1 - \frac{1}{W_\theta(t)}\right)^{k-1}.$$

When  $\alpha > \theta$  (resp.  $\alpha = \theta$ ,  $\alpha < \theta$ ), this new tree is supercritical (resp. critical, sub-critical) and we talk about *clonal supercritical case* (resp. *critical*, *sub-critical case*).

Moreover, the law of  $Z_0$  conditionally on the event  $\{N_t > 0\}$  can be obtained, and is given by

$$\mathbb{P}(Z_0(t) = k \mid N_t > 0) = \frac{e^{-\theta t} W(t)}{W_\theta(t)^2} \left(1 - \frac{1}{W_\theta(t)}\right)^{k-1}, \quad \forall k \geq 1. \quad (2.8)$$

For the rest of this paper, unless otherwise stated, the notation  $\mathbb{P}_t$  refers to  $\mathbb{P}(\cdot \mid N_t > 0)$  whereas  $\mathbb{P}_\infty$  refers to the probability measure conditioned on the non-extinction event (which has positive probability in the supercritical case).

Finally, we recall the asymptotic behavior of the scale functions  $W(t)$  and  $W_\theta(t)$  which is widely used in the sequel,

**Lemma 2.2.** ([6, Thm. 3.21]) *There exist a positive constant  $\gamma$  such that,*

$$e^{-\alpha t} \psi'(\alpha) W(t) - 1 = \mathcal{O}(e^{-\gamma t}).$$

*In the case that  $\theta < \alpha$  (clonal supercritical case),*

$$W_\theta(t) \underset{t \rightarrow \infty}{\sim} \frac{e^{(\alpha-\theta)t}}{\psi_\theta(\alpha-\theta)}.$$

*In the case that  $\theta > \alpha$  (clonal sub-critical case),*

$$W_\theta(t) = \frac{\theta}{\psi(\theta)} + \mathcal{O}(e^{-(\theta-\alpha)t}).$$

*In the case where  $\theta = \alpha$  (clonal critical case),*

$$W_\theta(t) \underset{t \rightarrow \infty}{\sim} \frac{\theta t}{\psi'(\alpha)}.$$

For a purpose, a more precise description of the asymptotic behavior of  $W$  is needed. It is given by the following result.

**Lemma 2.3.** [13, Prop. 5.1] *There exists a positive non-increasing càdlàg function  $F$  such that*

$$W(t) = \frac{e^{\alpha t}}{\psi'(\alpha)} - e^{\alpha t} F(t), \quad t \geq 0,$$

and

$$\lim_{t \rightarrow \infty} e^{\alpha t} F(t) = \begin{cases} \frac{1}{b\mathbb{E}V-1} & \text{if } \mathbb{E}V < \infty, \\ 0 & \text{otherwise.} \end{cases}$$

From this Lemma and (2.7), one can easily deduce that

$$\mathbb{P}(\text{NonEx}) = \lim_{t \rightarrow \infty} \mathbb{P}(N_t > 0) = \frac{\alpha}{b}, \quad (2.9)$$

where NonEx refer to the non-extinction event.

In [5], we show that a CPP stopped at time  $t$  with scale function  $W$  can be constructed by grafting independent CPP stopped at a fixed time  $a \leq t$  on a CPP stopped at time  $t - a$  with an explicit scale function different of  $W$  (see Figure 2). Moreover, we showed that the frequency spectrum can be expressed as an integral with respect to the random measure  $\mathcal{N}$  along the CPP, that is

$$\prod_{i=1}^l A(k_i, t) = \sum_{i=1}^l \int_{[0, t] \times \mathbb{N}} \mathbb{1}_{Z_0^{(u)}(a) = k_i} \sum_{u_{1:l-1}=1}^{N_{t-a}^{(t)}} \prod_{\substack{j=1 \\ i \neq j}}^{l-1} A^{(u_j)}(k_j, a) \mathcal{N}(da, du), \quad (2.10)$$

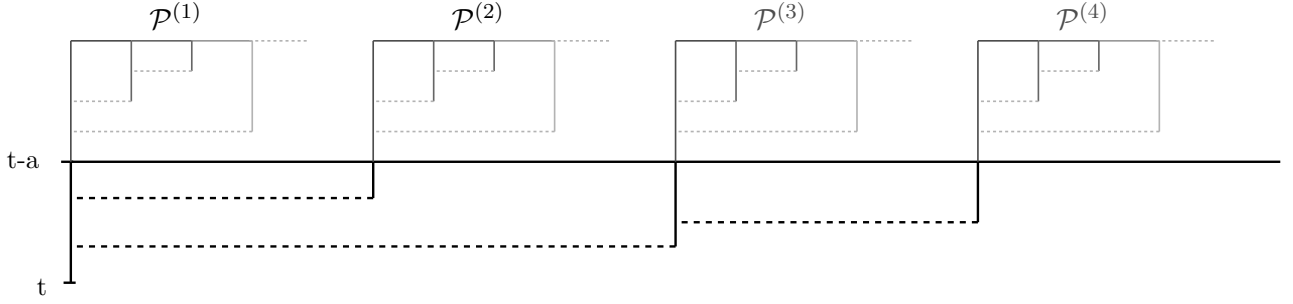


Figure 2: Adjunction of trees.

where  $A^{(u)}(k, a)$  (resp.  $Z_0^{(u)}$ ) refers to the frequency spectrum (resp. clonal family) of the  $u$ th grafted sub-CPP, and  $\sum_{u_1:l=1}^{N_{t-a}^{(t)}}$  denotes for the multi-sum

$$\sum_{u_1=1}^{N_{t-a}^{(t)}} \cdots \sum_{u_{l-1}=1}^{N_{t-a}^{(t)}}.$$

Moreover, in [5, Thm, 3.1] we show that the expectation of such integral can be computed easily when the integrand presents local independence properties with the random measure as in formula (2.10). Equation (2.10) is used later to obtain some moments estimates useful to prove our theorems. In particular, this allows to prove that (see [5]) for any positive integer  $k$  and  $l$ ,

$$\mathbb{E}_t A(k, t) = W(t) \int_0^t \frac{\theta e^{-\theta s}}{W_\theta(s)^2} \left(1 - \frac{1}{W_\theta(s)}\right)^{k-1} ds, \quad (2.11)$$

and

$$\begin{aligned} \mathbb{E} A(k, t) A(l, t) &= 2W(t)^2 \int_0^t \frac{\theta e^{-\theta s}}{W_\theta(s)^2} \left(1 - \frac{1}{W_\theta(s)}\right)^{k-1} ds \int_0^t \frac{\theta e^{-\theta s}}{W_\theta(s)^2} \left(1 - \frac{1}{W_\theta(s)}\right)^{l-1} ds \\ &\quad - W(t) \int_0^t 2\theta \frac{e^{-\theta a} W(a)}{W_\theta(a)^2} \left(1 - \frac{1}{W_\theta(a)}\right)^{l-1} \int_0^s \frac{\theta e^{-\theta s}}{W_\theta(s)^2} \left(1 - \frac{1}{W_\theta(a)}\right)^{k-1} ds da \\ &\quad - W(t) \int_0^t 2\theta \frac{e^{-\theta a} W(a)}{W_\theta(a)^2} \left(1 - \frac{1}{W_\theta(a)}\right)^{k-1} \int_0^s \frac{\theta e^{-\theta s}}{W_\theta(s)^2} \left(1 - \frac{1}{W_\theta(a)}\right)^{l-1} ds da \\ &\quad + W(t) \mathbb{E} \int_0^t \theta W(a)^{-1} (\mathbb{E} [A(k, t) \mathbb{1}_{Z_0(a)=l}] + \mathbb{E} [A(l, t) \mathbb{1}_{Z_0(a)=k}]) da \\ &\quad + \mathbb{1}_{l=k} W(t) \int_0^t \frac{\theta e^{-\theta s}}{W_\theta(s)^2} \left(1 - \frac{1}{W_\theta(s)}\right)^{k-1} ds. \end{aligned} \quad (2.12)$$

These tools also allow, for instance, to prove next two results [14, 8, 5].

**Theorem 2.4.** *There exists a random variable  $\mathcal{E}$ , such that*

$$\lim_{t \rightarrow \infty} e^{-\alpha t} N_t = \frac{\mathcal{E}}{\psi'(\alpha)}, \quad \text{a.s. and in } L^2.$$



Moreover, under  $\mathbb{P}_\infty$ ,  $\mathcal{E}$  is exponentially distributed with parameter one.

**Theorem 2.5.** *For any positive integer  $k$ ,*

$$\lim_{t \rightarrow \infty} e^{-\alpha t} A(k, t) = \frac{c_k \mathcal{E}}{\psi'(\alpha)}, \quad \text{a.s. and in } L^2,$$

where  $\mathcal{E}$  is the random variable of the Theorem 2.4 and

$$c_k = \int_0^\infty \frac{\theta e^{-\theta a}}{W_\theta(a)} \left(1 - \frac{1}{W_\theta(a)}\right)^{k-1} da. \quad (2.13)$$

### 3 Main results

The a.s. convergence stated in Section 2 suggests studying the second order properties of the convergence to get central limit theorem. Our main result, Theorem 3.5, allows to study the asymptotic error in the approximation 2 proposed in the introduction of this work. In addition, we prove more standard central limit theorems which are interesting from the theoretical point of view.

Before going further, we recall that the Laplace distribution with mean  $\mu \in \mathbb{R}^n$  and covariance matrix  $K$  is the probability distribution whose characteristic function is given, for all  $\lambda \in \mathbb{R}^n$  by

$$\frac{1}{1 + \frac{1}{2} \lambda' K \lambda - i \mu' \lambda}$$

We denote this law by  $\mathcal{L}(\mu, K)$ . We also recall that, if  $G$  is a Gaussian random vector with mean  $\mu$  and covariance matrix  $K$  and  $\mathcal{E}$  is an exponential random variable with parameter 1 independent of  $G$ , then  $\sqrt{\mathcal{E}}G$  is Laplace  $\mathcal{L}(\mu, K)$ .

#### 3.1 CLT for the convergence of Theorem 2.5

**Theorem 3.1.** *Suppose that  $\theta > \alpha$  and  $\int_{[0, \infty)} e^{(\theta - \alpha)v} \mathbb{P}_V(dv) > 1$ . Then, we have, under  $\mathbb{P}_\infty$ ,*

$$\left( e^{\alpha \frac{t}{2}} \left( \psi'(\alpha) A(k, t) - e^{\alpha t} c_k \mathcal{E} \right) \right)_{k \in \mathbb{N}} \xrightarrow[t \rightarrow \infty]{(d)} \mathcal{L}(0, K),$$

where  $K$  is some covariance matrix and the constants  $c_k$  are defined in (2.13).

The proof of this result can be found in Section 4.

**Remark 3.2.** *We are not able to compute explicitly the covariance matrix  $K$  in the general case due to our method of demonstration. However, all our other results give explicit formulas. In particular, the case where  $\mathbb{P}_V$  is exponential is given by the next theorem. The Yule case is also covered in the following theorem for  $d = 0$  although it does not satisfy the hypothesis of Theorem 3.1.*

**Theorem 3.3.** *Suppose that  $V$  is exponentially distributed with parameter  $d \in [0, b)$ . In this case,  $\alpha = b - d$ . We still suppose that  $\alpha < \theta$ , then*

$$\left( e^{\alpha \frac{t}{2}} \left( \psi'(\alpha) A(k, t) - e^{\alpha t} c_k \mathcal{E} \right) \right)_{k \in \mathbb{N}} \xrightarrow[t \rightarrow \infty]{(d)} \mathcal{L}(0, K), \quad \text{w.r.t. } \mathbb{P}_\infty,$$

where  $K$  is given by

$$K_{l,k} = M_{l,k} + c_k c_l \frac{\alpha}{b} \left(1 - 6 \frac{d}{\alpha}\right),$$

and

$$\begin{aligned} M_{l,k} = & 2\psi'(\alpha) \int_0^\infty \frac{\theta e^{-\theta a}}{W_\theta(a)^2} \left( \left(1 - \frac{1}{W_\theta(a)}\right)^{l-1} (\mathbb{E}_a[A(k, a)] - c_k W(a)) + \left(1 - \frac{1}{W_\theta(a)}\right)^{k-1} (\mathbb{E}_a[A(l, a)] - c_l W(a)) \right) da \\ & - \psi'(\alpha) \int_0^\infty \theta W(a)^{-1} \mathbb{E}_a[(A(k, a) - c_k N_a) \mathbb{1}_{Z_0(a)=l} + (A(l, a) - c_l N_a) \mathbb{1}_{Z_0(a)=k}] \\ & + \mathbb{1}_{l=k} \int_0^\infty \frac{\theta e^{-\theta s}}{W_\theta(s)^2} \left(1 - \frac{1}{W_\theta(s)}\right)^{k-1} ds, \end{aligned} \quad (3.1)$$

where  $W$ ,  $W_\theta$ ,  $\psi'(\alpha)$  are defined in the Section 2.

The proof of this result can be found in Section 6. Note that an explicit formula for  $\mathbb{E}_t A(k, t)$  is given by (2.11). Explicit formulas for  $\mathbb{E}_t [A(k, t) \mathbb{1}_{Z_0(t)=l}]$  can also be found in Proposition 4.5 of [5], and a formula for  $\mathbb{E}_t [N_a \mathbb{1}_{Z_0(t)=k}]$  can be found in Proposition 4.1 of [6].

**Remark 3.4.** *The condition on  $V$  in Theorem 3.1 is required only to ensure controls of the moments of the considered quantities. However, although the Yule case does not satisfy this condition ( $V = \infty$  p.s.) it is included in this last theorem ( $d=0$ ). This suggests that the condition on  $V$  may not be needed.*

### 3.2 CLT for the error between $A(k, t)$ and $c_k N_t$

The next theorem concerns the error between  $A(k, t)$  and  $c_k N_t$ . Once again, we have an explicit expression of the covariance matrix of the limit.

**Theorem 3.5.** *Suppose that  $\theta > \alpha$ , then*

$$\psi'(\alpha) \left( e^{\alpha \frac{t}{2}} (A(k, t) - c_k N_t) \right)_{k \in \mathbb{N}} \xrightarrow[t \rightarrow \infty]{(d)} \mathcal{L}(0, M), \text{ w.r.t. } \mathbb{P}_\infty,$$

where  $M$  is defined in relation (3.1).

The proof of this result can be found in Section 5.

**Remark 3.6.** *We do not know yet if the exponential random variable appearing the Gaussian mixing leading to a Laplace distribution is the same as the exponential limit of  $e^{-\alpha t} A(k, t)$ . However, the CLT for Markov branching processes in [3] suggest that it is, actually, the case. If, this is true in our case, it would be enough to know the correlations between the limits involved in Theorem 3.1 and 3.5 to obtain an explicit expression for the covariance matrix in Theorem 3.1.*

## 4 Proof of Theorem 3.1

The proof of this theorem is based on the proof of the central limit theorem for the process  $(N_t, t \in \mathbb{R}_+)$  given in [13]. The structure of the proof follows the same lines and is detailed in Section 4 of [13]. In a sake of conciseness, we only highlight the difficulties arising in our new context. The results which are straightforward rewording of the proofs given in [13] are left to the reader. However, we think it is necessary to recall some aspects of [13], in particular from [13, Section 4]. First, we recall that there exists a family  $(N_t^{(i)}, t \in \mathbb{R}_+)_{i \geq 1}$  of i.i.d. population counting processes with the same law as  $(N_t, t \in \mathbb{R}_+)$ , and a Poisson random measure  $\xi$  on  $\mathbb{R}_+$  with intensity  $b da$  such that

$$N_t = \int_{[0,t]} N_{t-u}^{(\xi_u)} \mathbf{1}_{V_\emptyset > u} \xi(du) + \mathbf{1}_{V_\emptyset > t}, \quad \text{almost surely,} \quad (4.1)$$

where  $\xi_u = \xi([0, u])$ . In addition, we have that  $t \rightarrow \mathbb{E}[N_t \mathcal{E}]$  is the unique solution bounded on finite intervals of the renewal equation,

$$\begin{aligned} f(t) &= \int_{\mathbb{R}_+} f(t-u) b e^{-\alpha u} \mathbb{P}(V > u) du \\ &\quad + \alpha b \mathbb{E}[N] \star \left( \int_{\mathbb{R}_+} e^{-\alpha v} \mathbb{P}(V > \cdot, V > v) dv \right) (t) \\ &\quad + \alpha \int_{\mathbb{R}_+} e^{-\alpha v} \mathbb{P}(V > t, V > v) dv, \end{aligned} \quad (4.2)$$

and it is given by

$$\mathbb{E}[N_t \mathcal{E}] = \left(1 + \frac{\alpha}{b} - e^{-\alpha t}\right) W(t) - (1 - e^{-\alpha t}) W \star \mathbb{P}_V(t). \quad (4.3)$$

We also recall that equation (4.2) is obtained by taking the product  $N_t N_s$ , for some real number  $t$  and  $s$ . Now, equation (4.1) allows to obtain a renewal equation for  $\mathbb{E}[N_t N_s]$  which leads to (4.2) when taking the limit in  $s$  of the renormalized equation. We also recall that Lemma 2.3 and equation (2.7) gives

$$\frac{1}{\mathbb{P}(N_t > 0)} = \frac{b}{\alpha} - \frac{b\mu\psi'(\alpha)}{\alpha} e^{-\alpha t} + o(e^{-\alpha t}). \quad (4.4)$$

This also leads, in conjunction with equation (4.3), to

$$\mathbb{E}_t N_t \mathcal{E} = \frac{2e^{\alpha t}}{\psi'(\alpha)} - \frac{1}{\psi'(\alpha)} - 3\mu + o(1). \quad (4.5)$$

Finally, let us recall that for any fixed time  $u$ , there is a natural order (for instance given by the contour process [14]) of the individuals alive at this time. Moreover, we denote, for  $1 \leq i \leq N_t$ ,  $O_i^{(u)}$  the residual lifetime of the  $i$ th individual alive at time  $u$ . The law of the vector  $(O_2^{(u)}, \dots, O_{N_u}^{(u)})$  is given by the following lemma which comes from [13].

**Lemma 4.1.** *Let  $u$  in  $\mathbb{R}_+$ , we denote by  $O_i$  for  $i$  an integer between 1 and  $N_u$  the residual lifetime of the  $i$ th individuals alive at time  $u$ . Then under  $\mathbb{P}_u$ , the family  $(O_i, i \in \{1, \dots, N_u\})$  form a family of independent random variables, independent of  $N_u$ , and, except  $O_1$ , having the same distribution, given by, for  $2 \leq i \leq N_u$ ,*

$$\mathbb{P}_u(O_i \in dx) = \int_{\mathbb{R}_+} \frac{W(u-y)}{W(u)-1} b\mathbb{P}(V-y \in dx) dy. \quad (4.6)$$

*Moreover, it follows that the family  $(N_s(O_i), s \in \mathbb{R}_+)_{1 \leq i \leq N_u}$  is an independent family of process, i.i.d. for  $i \geq 2$ , and independent of  $N_u$ .*

To end this reminder, let us recall the decomposition of the limiting random variable  $\mathcal{E}$  (given for instance in Theorem 2.4) at a fixed time  $u$ .

**Lemma 4.2.** *[13, Lemma 6.8] We have the following decomposition of  $\mathcal{E}$ ,*

$$\mathcal{E} = e^{-\alpha u} \sum_{i=1}^{N_u} \mathcal{E}_i(O_i), \quad a.s.$$

*Moreover, under  $\mathbb{P}_u$ , the random variables  $(\mathcal{E}_i(O_i))_{i \geq 1}$  are independent, independent of  $N_u$ , and identically distributed for  $i \geq 2$ .*

We can now start the proof of theorem 3.1. As in [13], the proof begins by some estimate on moments.

## 4.1 Preliminary moments estimates

We start by computing the moment in the case of a standard splitting tree. According to [13, Section 4], the next step is to obtain the same kind of estimates in the case of a splitting tree whose ancestor individual has a lifetime distribution which can be different from the rest of the population.

### 4.1.1 Case $V_\emptyset \stackrel{\mathcal{L}}{=} V$

One of the main difficulties to extend the preceding proof to the frequency spectrum is to get estimates on

$$\mathbb{E} \left[ (\psi'(\alpha) A(k, t) - e^{\alpha t} c_k \mathcal{E})^n \right], \text{ for } n = 2 \text{ or } 3.$$

We first study the renewal equation satisfied by  $\mathbb{E} A(k, t) \mathcal{E}$  similarly as in [13, Lemma 6.1].

**Lemma 4.3** (Joint moment of  $\mathcal{E}$  and  $A(k, t)$ ).  *$\mathbb{E} [A(k, t) \mathcal{E}]$  is the unique solution bounded on finite intervals of the renewal equation,*

$$\begin{aligned} f(t) &= \int_{\mathbb{R}_+} f(t-u) b e^{-\alpha u} \mathbb{P}(V > u) du \\ &\quad + \alpha \mathbb{E} [A(k, \cdot)] \star b \left( \int_{\mathbb{R}_+} e^{-\alpha v} \mathbb{P}(V > \cdot, V > v) dv \right) (t) \\ &\quad + \alpha \mathbb{E} [\mathcal{E} X_t], \end{aligned} \quad (4.7)$$

with  $X_t$  the number of families of size  $k$  alive at time  $t$  whose original mutation has taken place during the lifetime of the ancestor individual.

*Proof.* We recall that  $A(k, t)$  is the number of non-ancestral families of size  $k$  at time  $t$ . Similarly, as for  $N_t$ ,  $A(k, t)$  can be obtained as the sum of the contributions of all the trees grafted on the lifetime of the ancestor individual in addition to the mutations which take place on the ancestral branch, that is,

$$A(k, t) = \int_{[0, t]} A(k, t - u, \xi_u) \mathbb{1}_{V_\emptyset > u} \xi(du) + X_t,$$

where  $(A(k, t, i), t \in \mathbb{R}_+)_{i \geq 1}$  is a family of independent processes having the same law as  $A(k, t)$ . Now, taking the product  $A(k, t)N_s$  and using the same arguments as in the proof of lemma [13, Lemma 6.1] to take the limit in  $s$  leads to the result. In particular, the last term is obtained using that

$$\lim_{s \rightarrow \infty} \mathbb{E} \left[ X_t \frac{N_s}{W(s)} \right] = \mathbb{E} [X_t \mathcal{E}].$$

□

The result of Lemma 4.3 is quite disappointing since the presence of the mysterious process  $X_t$  prevents any explicit resolution of equation (4.7). However, one may note that equation (4.7) is quite similar to equation (4.2) driving  $\mathbb{E}N_t \mathcal{E}$ , so if the contribution of  $X_t$  in the renewal structure of the process is small enough, one can expect the same asymptotic behavior for  $\mathbb{E}A(k, t) \mathcal{E}$  as for  $\mathbb{E}N_t \mathcal{E}$ . Moreover, we clearly have on  $X_t$  the following a.s. estimate,

$$X_t \leq \int_{[0, t]} \mathbb{1}_{Z_0^{(u)}(t-u) > 0} \mathbb{1}_{V > u} \xi(du), \quad (4.8)$$

where  $Z_0^{(i)}$  denote for the ancestral families on the  $i$ th trees grafted on the ancestral branch. Hence, if we take  $\theta > \alpha$  and we suppose  $V < \infty$  a.s., one can expect that  $X_t$  decreases very fast. These are the ideas the following Lemma is based on. Moreover, as it is seen in the proof of the following lemma, the hypothesis  $V < \infty$  a.s. can be weakened.

**Lemma 4.4.** *Under the hypothesis of Theorem 3.1, for all  $k \geq 1$ , there exists a constant  $\gamma_k \in \mathbb{R}$  such that,*

$$\lim_{t \rightarrow \infty} \mathbb{E}N_t \mathcal{E} c_k - \mathbb{E}A(k, t) \mathcal{E} = \gamma_k. \quad (4.9)$$

*Proof.* Combining equations (4.2) and (4.7), we get that,

$$\begin{aligned} \mathbb{E}N_t \mathcal{E} c_k - \mathbb{E}A(k, t) \mathcal{E} &= \int_{\mathbb{R}_+} (\mathbb{E}N_{t-u} \mathcal{E} c_k - \mathbb{E}A(k, t-u) \mathcal{E}) b e^{-\alpha u} \mathbb{P}(V > u) du \\ &\quad + \underbrace{\alpha b (c_k \mathbb{E}N - \mathbb{E}[A(k, \cdot)]) \star \left( \int_{\mathbb{R}_+} e^{-\alpha v} \mathbb{P}(V > \cdot, V > v) dv \right)}_{:= \xi_1^{(k)}(t)}(t) \\ &\quad + \underbrace{c_k \mathbb{P}(V > t) - \alpha \mathbb{E}[X_t \mathcal{E}]}_{:= \xi_2^{(k)}(t)}, \end{aligned}$$

which is also a renewal equation. On one hand, using equations (2.5) and (2.11) imply that

$$\mathbb{E}_t [c_k N_t - A(k, t)] = W(t) \int_t^\infty \frac{\theta e^{-\theta s}}{W_\theta(s)^2} \left(1 - \frac{1}{W_\theta(s)}\right)^{k-1} ds,$$

which leads using Lemma 2.2, to

$$\begin{aligned} \xi_1(t) &= \alpha \int_{\mathbb{R}_+} (c_k \mathbb{E} N_{t-u} - \mathbb{E} [A(k, t-u)]) \int_{\mathbb{R}_+} e^{-\alpha v} \mathbb{P}(V > u, V > v) dv du \\ &\leq \mathcal{C} \int_{[0,t]} e^{(\alpha-\theta)t-u} \mathbb{P}(V > u) du \int_{[0,\infty)} e^{-\alpha u} du \\ &\leq \frac{\mathcal{C}}{\alpha} e^{-(\theta-\alpha)t} \int_0^t e^{(\theta-\alpha)u} \mathbb{P}(V > u) du, \end{aligned} \quad (4.10)$$

for some positive real constant  $\mathcal{C}$ .

The derivative of the r.h.s. of (4.10) is given by

$$\frac{\mathcal{C}}{\alpha} e^{-(\theta-\alpha)t} \left( e^{(\theta-\alpha)t} \mathbb{P}(V > t) - (\alpha - \theta) \int_0^t e^{(\theta-\alpha)u} \mathbb{P}(V > u) du \right), \quad t > 0, \quad (4.11)$$

which is equal to

$$\frac{\mathcal{C}}{\alpha} e^{-(\theta-\alpha)t} \left( 1 - \int_{[0,t]} e^{(\theta-\alpha)s} \mathbb{P}_V(ds) \right), \quad t > 0,$$

using Stieljes integration by parts. Now, since,

$$\int_{[0,\infty)} e^{(\theta-\alpha)s} \mathbb{P}_V(ds) > 1,$$

this shows that the right hand side of (4.10) is decreasing for  $t$  large enough. Moreover, it is straightforward to show that the r.h.s. of (4.10) is also integrable. This implies that  $\xi_1^{(k)}$  is DRI from the same Lemma. On the other hand, it follows from (4.8) that

$$X_t \mathcal{E} \leq \mathcal{E} \int_{[0,t]} \mathbb{1}_{Z_0^{(u)}(t-u) > 0} \mathbb{1}_{V > t} \xi(du). \quad (4.12)$$

Then, we obtain using Cauchy-Schwarz inequality, that

$$\mathbb{E} [X_t \mathcal{E}] \leq \sqrt{\frac{2\alpha}{b}} \mathbb{E} \left[ \left( \int_{[0,t]} \mathbb{1}_{Z_0^{(u)}(t-u) > 0} \mathbb{1}_{V > t} \xi(du) \right)^2 \right]^{1/2}.$$

It follows that we need to investigate the behavior of

$$\mathbb{E} \left[ \left( \int_{(0,t)} \mathbb{1}_{Z_0^{(u)}(t-u) > 0} \mathbb{1}_{V > t} \xi(du) \right)^2 \right],$$

which is equal to

$$\int_0^t \mathbb{P}(Z_0(t-u) > 0) \mathbb{P}(V > t) b du + \int_{[0,t]^2} \mathbb{P}(Z_0(t-v) > 0) \mathbb{P}(Z_0(t-u) > 0) \mathbb{P}(V > u, V > v) b^2 du dv,$$

using [13, Lemma 2.6]. Then, since, from (2.8) and Lemma 2.2,

$$\mathbb{P}_{t-u}(Z_0(t-u) > 0) = \frac{e^{-\theta(t-u)}W(t-u)}{W_\theta(t-u)} = \mathcal{O}(e^{-(\theta-\alpha)(t-u)}),$$

it follows, using that the right hand side of (4.10) is DRI and Lemma A.1, that  $\xi_2^{(k)}$  is DRI. Finally, it comes from Theorem A.2, that

$$\lim_{t \rightarrow \infty} \mathbb{E} N_t \mathcal{E} c_k - \mathbb{E} A(k, t) \mathcal{E} = \frac{\alpha}{\psi'(\alpha)} \int_{\mathbb{R}_+} \xi_1^{(k)}(s) + \xi_2^{(k)}(s) ds. \quad (4.13)$$

□

Using the preceding lemma, we can now get the quadratic error in the convergence of the frequency spectrum.

**Lemma 4.5** (Quadratic error for the convergence of  $A(k, t)$ ). *Let  $k$  and  $l$  two positive integers. Then under the hypothesis of Theorem 3.1, there exists a family of real numbers  $(a_{k,l})_{l,k \geq 1}$  such that,*

$$\lim_{t \rightarrow \infty} e^{-\alpha t} \mathbb{E} [(\psi'(\alpha)A(k, t) - e^{\alpha t} \mathcal{E} c_k) (\psi'(\alpha)A(l, t) - e^{\alpha t} \mathcal{E} c_l)] = \frac{\alpha}{b} a_{k,l},$$

where the sequence  $(c_k)_{k \geq 1}$  is defined by (2.13).

*Proof.* Now, noting

$$c_k(t) := \int_0^t \frac{\theta e^{-\theta a}}{W_\theta(a)^2} \left(1 - \frac{1}{W_\theta(a)}\right)^{k-1} da, \quad (4.14)$$

we have, from (2.12) and Lemma 2.3,

$$\psi'(\alpha)^2 \mathbb{E}_t [A(k, t)A(l, t)] = 2e^{2\alpha t} c_k(t) c_l(t) + e^{\alpha t} \left( 4\psi'(\alpha) e^{\alpha t} F(t) c_k(t) c_l(t) + \frac{R}{\psi'(\alpha)} \right) + \mathcal{O}(1), \quad (4.15)$$

with

$$\begin{aligned} R := & -\psi'(\alpha) \int_0^\infty 2\theta \frac{e^{-\theta a} W(a)}{W_\theta(a)^2} \left(1 - \frac{1}{W_\theta(a)}\right)^{l-1} \int_0^a \frac{e^{-\theta s}}{W_\theta(s)^2} \left(1 - \frac{1}{W_\theta(a)}\right)^{k-1} ds da \\ & - \psi'(\alpha) \int_0^\infty 2\theta \frac{e^{-\theta a} W(a)}{W_\theta(a)^2} \left(1 - \frac{1}{W_\theta(a)}\right)^{k-1} \int_0^a \frac{e^{-\theta s}}{W_\theta(s)^2} \left(1 - \frac{1}{W_\theta(a)}\right)^{l-1} ds da \\ & + \psi'(\alpha) \int_0^\infty \theta W(a)^{-1} (\mathbb{E}_t [A(k, t) \mathbb{1}_{Z_0(a)=l}] + \mathbb{E}_t [A(l, t) \mathbb{1}_{Z_0(a)=k}]) da, \end{aligned}$$

and  $F, \mu$  are defined in Lemma 2.3. Now, using (4.4), we have

$$\mathbb{E}_t \mathcal{E}^2 - 2 = -2\mu \psi'(\alpha) e^{-\alpha t} + o(e^{-\alpha t}),$$

which leads to

$$\begin{aligned}
& \mathbb{E}_t \left[ (e^{-\alpha t} \psi'(\alpha) A(k, t) - \mathcal{E} c_k) (e^{-\alpha t} \psi'(\alpha) A(l, t) - \mathcal{E} c_l) \right] \\
&= \mathbb{E}_t \left[ e^{-2\alpha t} \psi'(\alpha)^2 A(k, t) A(l, t) \right] - c_l \mathbb{E}_t \left[ e^{-\alpha t} \psi'(\alpha) A(k, t) \mathcal{E} \right] - c_k \mathbb{E}_t \left[ e^{-\alpha t} \psi'(\alpha) A(l, t) \mathcal{E} \right] \\
&\quad + 2c_k c_l - 2c_k c_l \mu \psi'(\alpha) e^{-\alpha t} + o(e^{-\alpha t}), \\
&= 2(c_k(t) - c_k)(c_l(t) - c_l) - 4\mu \psi'(\alpha) c_k c_l e^{-\alpha t} + R e^{-\alpha t} \\
&\quad - (2c_k(t) c_l + 2c_l(t) c_k - 2c_k c_l \psi'(\alpha) e^{-\alpha t} \mathbb{E}_t N_t \mathcal{E}) \\
&\quad + \psi'(\alpha) c_l e^{-\alpha t} \mathbb{E}_t [(c_k N_t - A(k, t)) \mathcal{E}] + \psi'(\alpha) c_k e^{-\alpha t} \mathbb{E}_t [(c_l N_t - A(l, t)) \mathcal{E}] + o(e^{-\alpha t}),
\end{aligned}$$

Since, by Lemma 2.2

$$c_k(t) = c_k + \mathcal{O}(e^{-\theta t}) = c_k + o(e^{-\alpha t}),$$

it follows, combining (4.13), (4.5), and Lemma 4.4, that

$$\begin{aligned}
& e^{\alpha t} \mathbb{E}_t \left[ (e^{-\alpha t} \psi'(\alpha) A(k, t) - \mathcal{E} c_k) (e^{-\alpha t} \psi'(\alpha) A(l, t) - \mathcal{E} c_l) \right] \\
&= \psi'(\alpha) (c_k \gamma_l + c_l \gamma_k) + c_k c_l (2e^{\alpha t} - 2\psi'(\alpha) \mathbb{E}_t N_t \mathcal{E}) + R - 4\mu \psi'(\alpha) c_k c_l + o(1) \\
&= \psi'(\alpha) (c_k \gamma_l + c_l \gamma_k) + c_k c_l \left( \frac{1}{\psi'(\alpha)} + 3\mu \right) + R - 4\mu \psi'(\alpha) c_k c_l + o(1).
\end{aligned}$$

The result follows readily from the fact that  $\mathbb{P}(N_t > 0) \sim \frac{\alpha}{b}$ .

□

**Lemma 4.6** (Boundedness of the third moment). *Let  $k_1, k_2, k_3$  three positive integers, then*

$$\mathbb{E} \left[ \prod_{i=1}^3 \left| e^{-\frac{\alpha}{2} t} (\psi'(\alpha) A(k_i, t) - e^{\alpha t} \mathcal{E} c_{k_i}) \right| \right] = \mathcal{O}(1).$$

*Proof.* We have,

$$\mathbb{E} \left[ \left| \prod_{i=1}^3 \frac{(\psi'(\alpha) A(k_i, t) - e^{\alpha t} \mathcal{E} c_{k_i})}{e^{\frac{\alpha}{2} t}} \right| \right] \leq \prod_{i=1}^3 \left( \mathbb{E} \left[ \left| \frac{(\psi'(\alpha) A(k_i, t) - e^{\alpha t} \mathcal{E} c_{k_i})}{e^{\frac{\alpha}{2} t}} \right|^3 \right] \right)^{\frac{1}{3}}.$$

Hence, we only have to prove the Lemma for  $k_1 = k_2 = k_3 = k$ . Hence,

$$\begin{aligned}
\mathbb{E} \left[ \left| \frac{(\psi'(\alpha) A(k, t) - e^{\alpha t} \mathcal{E} c_k)}{e^{\frac{\alpha}{2} t}} \right|^3 \right] &\leq 8 \mathbb{E} \left[ \left| \frac{\psi'(\alpha) A(k, t) - c_k N_t}{e^{\frac{\alpha}{2} t}} \right|^3 \right] + 8c_k \mathbb{E} \left[ \left| \frac{\psi'(\alpha) N_t - N_t^\infty}{e^{\frac{\alpha}{2} t}} \right|^3 \right] \\
&\quad + 8c_k \mathbb{E} \left[ \left| \frac{N_t^\infty - e^{\alpha t} \mathcal{E}}{e^{\frac{\alpha}{2} t}} \right|^3 \right].
\end{aligned}$$



The last two terms have been treated in the proof of [13, Lemma 6.4], and the boundedness of

$$\mathbb{E} \left[ \left| \frac{\psi'(\alpha)A(k, t) - c_k N_t}{e^{\frac{\alpha}{2}t}} \right|^3 \right],$$

follows from the following Lemma 4.7 and Hölder's inequality. □

**Lemma 4.7.** *For all  $k \geq 1$ ,*

$$\mathbb{E} \left[ \left( \frac{A(k, t) - c_k N_t}{e^{-\frac{\alpha}{2}t}} \right)^4 \right],$$

*is bounded.*

Due to technicality, the proof of this lemma is postponed to the end in appendix.

#### 4.1.2 Arbitrary initial distribution case

The following Lemmas are the counter part of Lemmas 6.5, 6.6, and 6.7 of [13]. They play the same role in the proof of Theorem 3.1 as in the proof of the central limit theorem given in [13]. In the sequel, we denote by  $(A(k, t, \Xi))_{k \geq 1}$ , the frequency spectrum of the splitting tree where the lifetime of the ancestral individual is  $\Xi$ , in the same manner as for  $N_t(\Xi)$  in [13].

$$\mathcal{E}_i := \lim_{t \rightarrow \infty} \psi'(\alpha) e^{-\alpha t} N_t^i, \quad a.s., \quad (4.16)$$

and, let  $\mathcal{E}(\Xi)$  be the random variable defined by

$$\mathcal{E}(\Xi) := \int_{[0, \infty]} \mathcal{E}(\xi_u) e^{-\alpha u} \mathbf{1}_{\Xi > u} \xi(du). \quad (4.17)$$

**Lemma 4.8** ( $L^2$  convergence in the general case). *Consider the general frequency spectrum  $(A(k, t, \Xi))_{k \geq 1}$ , then, for all  $k$ ,  $\psi'(\alpha) e^{-\alpha t} A(k, t, \Xi)$  converge to  $\mathcal{E}(\Xi)$  (see 4.17) in  $L^2$  as  $t$  goes to infinity and*

$$\lim_{t \rightarrow \infty} e^{-\alpha t} \mathbb{E} \left[ (\psi'(\alpha) A(k, t, \Xi) - e^{\alpha t} \mathcal{E}(\Xi) c_k) (\psi'(\alpha) A(l, t, \Xi) - e^{\alpha t} \mathcal{E}(\Xi) c_l) \right] = \frac{\alpha}{b} a_{k,l} \int_{\mathbb{R}_+} e^{-\alpha u} \mathbb{P}(\Xi > u) b du,$$

where the convergence is uniform w.r.t. the random variable  $\Xi$ . In the case where  $\Xi$  is distributed as  $O_2^{(\beta t)}$ , for  $0 < \beta < \frac{1}{2}$ , we get

$$\lim_{t \rightarrow \infty} e^{-\alpha t} \mathbb{E} \left[ \left( \psi'(\alpha) A(k, t, O_2^{(\beta t)}) - e^{\alpha t} \mathcal{E}(O_2^{(\beta t)}) c_k \right) \left( \psi'(\alpha) A(l, t, O_2^{(\beta t)}) - e^{\alpha t} \mathcal{E}(O_2^{(\beta t)}) c_l \right) \right] = \psi'(\alpha) a_{k,l}.$$

**Lemma 4.9** (First moment). *The first moments are asymptotically bounded, that is, for all  $k \geq 1$ ,*

$$\mathbb{E} \left( \psi'(\alpha) A(k, t)(\Xi) - e^{\alpha t} c_k \mathcal{E}(\Xi) \right) \leq \mathcal{O}(1),$$

*uniformly with respect to the random variable  $\Xi$ .*

**Lemma 4.10** (Boundedness in the general case.). *Let  $k_1, k_2, k_3$  three positive integers, then*

$$\mathbb{E} \left[ \left| \prod_{i=1}^3 \frac{(\psi'(\alpha)A(k_i, t) - e^{\alpha t} \mathcal{E} c_{k_i})}{e^{\frac{\alpha}{2}t}} \right| \right] = \mathcal{O}(1),$$

*uniformly with respect to the random variable  $\Xi$ .*

We do not detail the proofs of these results since they are direct adaptations of the proofs of Lemmas 6.5, 6.6, and 6.7 of [13].

## 4.2 Proof of the result

The following result is based on the fact that, in the clonal sub-critical case, the lifetime of a family is expected to be small. It follows that one can expect that all the family of size  $k$  live in different subtrees as soon as  $t \gg u$ . This is the point of the following lemma.

**Lemma 4.11.** *Suppose that  $\alpha < \theta$ . If we denote by  $\Gamma_{u,t}$  the event,*

$$\Gamma_{u,t} = \{ \text{"there is no family in the population at time } t \text{ which is older than } u" \},$$

*then, for all  $\beta$  in  $(0, 1 - \frac{\alpha}{\theta})$ , we have*

$$\lim_{t \rightarrow \infty} \mathbb{P}_{\beta t}(\Gamma_{\beta t, t}) = 1.$$

*Proof.* The proof of this Lemma, as the calculation of the moments of  $A(k, t)$  relies on the representation of the genealogy of the living population at time  $t$  as a coalescent point process [5]. Moreover, we denote by  $\tilde{N}_u^{(t)}$  the number of living individuals at time  $u$  who have alive descent at time  $t$ . In [5], we showed that, under  $\mathbb{P}_t$ ,  $\tilde{N}_u^{(t)}$  is geometrically distributed with parameter  $\frac{W(t-u)}{W(t)}$ .

Now,  $\mathbb{1}_{\Gamma_{u,t}}$  can be rewritten as

$$\mathbb{1}_{\Gamma_{u,t}} = \prod_{i=1}^{\tilde{N}_u^{(t)}} \mathbb{1}_{\{Z_0^i(t-u)=0\}},$$

where  $Z_0^i(t-u)$  denotes the number of individuals alive at time  $t$  descending from the  $i$ th individual alive at time  $u$  and carrying its type (the clonal type of the sub-CPP). Moreover, from Proposition 4.3 of [5], we know that that under  $\mathbb{P}_t$ , the family  $Z_0^{(i)}(t-u)$  is an i.i.d. family of random variables distributed as  $Z_0(t-u)$  under  $\mathbb{P}_{t-u}$ , and  $\tilde{N}_u^{(t)}$  is independent of  $Z_0^{(i)}(t-u)$  (still under  $\mathbb{P}_t$ ).

Then,

$$\mathbb{P}_t(\Gamma_{t,u}) = \mathbb{E}_t \left[ \mathbb{P}_{t-u}(Z_0(t-u) = 0)^{\tilde{N}_u^{(t)}} \right] = \frac{\mathbb{P}_{t-u}(Z_0(t-u) = 0) \frac{W(t-u)}{W(t)}}{1 - \mathbb{P}_{t-u}(Z_0(t-u) = 0) \left(1 - \frac{W(t-u)}{W(t)}\right)}.$$

Using (2.8), some calculus leads to,

$$\mathbb{P}_t(\Gamma_{t,u}) = 1 - \frac{1}{1 + \frac{W_\theta(t-u)}{e^{-\theta(t-u)}W(t)} \left(1 - \frac{e^{-\theta(t-u)}W(t-u)}{W_\theta(t-u)}\right)}.$$

Now, since,

$$\mathbb{P}_t(\Gamma_{t,u}) = \mathbb{P}_u(\Gamma_{t,u}) \frac{\mathbb{P}(N_u > 0)}{\mathbb{P}(N_t > 0)} + \frac{\mathbb{P}(\Gamma_{t,u}, N_t = 0, N_u > 0)}{\mathbb{P}(N_t > 0)},$$

taking  $u = \beta t$ , we obtain, using Lemma 2.2 and

$$\mathbb{P}(N_t = 0, N_{\beta t} > 0) = \mathbb{P}(N_{\beta t} > 0) - \mathbb{P}(N_t > 0) \xrightarrow{t \rightarrow \infty} 0,$$

the desired result.  $\square$

*Proof of Theorem 3.1.* Fix  $0 < u < t$ . Note that the event  $\Gamma_{u,t}$  of Lemma 4.11 can be rewritten as

$$\mathbb{1}_{\Gamma_{u,t}} = \prod_{i=1}^{N_u} \mathbb{1}_{\{Z_0^i(t-u, O_i)=0\}}, \quad (4.18)$$

where  $Z_0^i(t-u, O_i)$  denote the number of individuals alive at time  $t$  carrying the same type as the  $i$ th alive individual at time  $u$ , that is the ancestral family of the splitting constructed from the residual lifetime of the  $i$ th individual (see Section 4 in [13]).

Let  $K$  be a multi-integer, we denote by  $\mathcal{L}^{(K)}$  (resp.  $A(K, t)$ ) the random vector  $(\mathcal{L}^{k_1}, \dots, \mathcal{L}^{k_N})$  (resp.  $(A(k_1, t), \dots, A(k_N, t))$ ) with

$$\mathcal{L}_t^{k_i} = \frac{\psi'(\alpha)A(k, t) - c_k e^{\alpha t} \mathcal{E}}{e^{\frac{\alpha}{2}t}}.$$

On the event  $\Gamma_{u,t}$ , we have a.s.,

$$A(k_l, t) = \sum_{i=1}^{N_u} A^{(i)}(k_l, t-u, O_i), \quad \forall l = 1, \dots, N,$$

where the family  $(A^{(i)}(k_l, t-u, O_i))_{i \geq 1}$  stand for the frequency spectrum for each subtree, which are independent from Lemma 4.1. Hence, using Lemma 4.2,

$$\mathcal{L}_t^{k_l} = \sum_{i=1}^{N_u} \frac{\psi'(\alpha)A^{(i)}(k_l, t-u, O_i) - e^{\alpha(t-u)} \mathcal{E}_i(O_i) c_{k_l}}{e^{\frac{\alpha}{2}u} e^{\frac{\alpha}{2}(t-u)}}.$$

By Lemma 4.1, that the family  $(A^{(i)}(k_l, t-u, O_i))_{2 \leq i \leq N_u}$  is i.i.d. under  $\mathbb{P}_u$ .

In the sequel, we denote, for all  $l$  and  $i \geq 1$ ,

$$\tilde{A}^{(i)}(k_l, t-u, O_i) = \frac{\psi'(\alpha)A^{(i)}(k_l, t-u, O_i) - e^{\alpha(t-u)} \mathcal{E}_i(O_i) c_{k_l}}{e^{\frac{\alpha}{2}(t-u)}}.$$

Now, let

$$\begin{aligned} \varphi_K(\xi) &:= \mathbb{E} \left[ \exp \left( i < \tilde{A}(K, t-u, O_2), \xi > \right) \mathbb{1}_{Z_0^2(t-u, O_2)=0} \right], \\ \tilde{\varphi}_K(\xi) &:= \mathbb{E} \left[ \exp \left( i < \tilde{A}(K, t-u, O_1), \xi > \right) \mathbb{1}_{Z_0^1(t-u, O_1)=0} \right]. \end{aligned}$$

From this point, following closely the proof of Theorem 3.2 of [13]. Taking  $u = \beta$  in  $(0, \frac{1}{2} \wedge (1 - \frac{\alpha}{\theta}))$ , the only difficulty is to handle the indicator function  $\mathbb{1}_{Z_0(t-u, O_i) > 0}$  in the Taylor development of  $\varphi_K$ . We show how it can be done for one of the second order terms, and leave the rest of the details to the reader.

It follows from Hlder's inequality that

$$\begin{aligned} & \mathbb{E} \left[ \left( \frac{\psi'(\alpha) A^{(i)}(k_l, (1-\beta)t, O_i) - e^{\alpha((1-\beta)t)} \mathcal{E}_i(O_i) c_{k_l}}{e^{\frac{\alpha}{2}((1-\beta)t)}} \right)^2 \mathbb{1}_{Z_0^2((1-\beta)t, O_2) > 0} \right] \\ & \leq \mathbb{E} \left[ \left( \frac{\psi'(\alpha) A^{(i)}(k_l, (1-\beta)t, O_i) - e^{\alpha(1-\beta)t} \mathcal{E}_i(O_i) c_{k_l}}{e^{\frac{\alpha}{2}(1-\beta)t}} \right)^3 \right]^{\frac{2}{3}} \mathbb{P}(Z_0^2((1-\beta)t, O_2) > 0)^{\frac{1}{3}}, \quad (4.19) \end{aligned}$$

from which it follows, using Lemma 4.10, that the r.h.s. of this last inequality is  $\mathcal{O}(\mathbb{P}(Z_0^2(t-u, O_2) > 0)^{\frac{1}{3}})$ . Now, using (4.18) and Lemma 4.11, it is easily seen that

$$\lim_{t \rightarrow \infty} \mathbb{P}(Z_0^2((1-\beta)t, O_2) > 0) = 0.$$

Finally, using Lemma 4.5, we get

$$\lim_{t \rightarrow \infty} \mathbb{E} \left[ \left( \frac{\psi'(\alpha) A^{(i)}(k_l, t-u, O_i) - e^{\alpha(t-u)} \mathcal{E}_i(O_i) c_{k_l}}{e^{\frac{\alpha}{2}(t-u)}} \right)^2 \mathbb{1}_{Z_0^2(t-u, O_2) = 0} \right] = \psi'(\alpha) a_{k,k}.$$

These allow us to conclude that

$$\lim_{t \rightarrow \infty} \mathbb{E}_{\beta t} \left[ e^{i < \mathcal{L}_t^{(K)}, \xi >} \mathbb{1}_{\Gamma_t} \right] = \frac{1}{1 + \sum_{i,j=1}^N \mathcal{M}_{i,j} \xi_i \xi_j},$$

where  $K_{i,j}$  is given by

$$\mathcal{M}_{i,j} := \psi'(\alpha) a_{K_i, K_j},$$

with  $K$  is the multi-integer  $(k_1, \dots, k_N)$ , and the  $a_{l,k}$ s are defined in Lemma 4.5.

To end the proof, note that,

$$\left| \mathbb{E}_{\infty} \left[ e^{i < \mathcal{L}_t^{(K)}, \xi >} \right] - \mathbb{E}_{\beta t} \left[ e^{i < \mathcal{L}_t^{(K)}, \xi >} \mathbb{1}_{\Gamma_{\beta t, t}} \right] \right| \leq \mathbb{E} \left[ \left| \frac{\mathbb{1}_{\text{NonEx}}}{\mathbb{P}(\text{NonEx})} - \frac{\mathbb{1}_{N_{\beta t} > 0} \mathbb{1}_{\Gamma_{\beta t, t}}}{\mathbb{P}(N_{\beta t} > 0)} \right| \right] \xrightarrow{t \rightarrow \infty} 0,$$

thanks to Lemma 4.11. □

## 5 Proof of Theorem 3.5

Since all the ideas of the proof of this theorem have been developed the preceding sections, we do not detail all the proof. The only step which needs clarification is the computation of the covariance

matrix of the Laplace limit law  $\mathcal{M}$ . According to the proof of Theorem 3.1, it is given by

$$\mathcal{M}_{i,j} := \lim_{t \rightarrow \infty} \frac{W(\beta t)}{e^{\alpha \beta t}} \mathbb{E} \left[ \left( \frac{\psi'(\alpha) A^{(i)}(k_i, (1-\beta)t, O_i) - \psi'(\alpha) c_{k_i} N_{(1-\beta)t}}{e^{\frac{\alpha}{2}((1-\beta)t)}} \right) \right. \\ \left. \times \left( \frac{\psi'(\alpha) A^{(i)}(k_j, (1-\beta)t, O_i) - c_{k_j} N_{(1-\beta)t}}{e^{\frac{\alpha}{2}((1-\beta)t)}} \right) \mathbb{1}_{Z_0^2((1-\beta)t, O_2) > 0} \right],$$

which is equal, thanks to (4.19) and an easy adaptation of Lemma 6.6 in [13], to

$$\mathcal{M}_{i,j} = \lim_{t \rightarrow \infty} \frac{b\psi'(\alpha)}{\alpha} \frac{W(\beta t)}{e^{\alpha \beta t}} e^{\alpha t} \mathbb{E} \left[ (e^{-\alpha t} A(k_i, t) - c_{k_i} e^{-\alpha t} N_t) (e^{-\alpha t} A(k_j, t) - c_{k_j} e^{-\alpha t} N_t) \right].$$

So it remains to get the limit of

$$e^{\alpha t} \mathbb{E} \left[ (e^{-\alpha t} \psi'(\alpha) A(k, t) - \psi'(\alpha) c_k e^{-\alpha t} N_t) (e^{-\alpha t} \psi'(\alpha) A(l, t) - c_l e^{-\alpha t} \psi'(\alpha) N_t) \right],$$

as  $t$  goes to infinity. We recall that using the calculus made in the proof of Theorem 6.3 of [5], we have

$$\mathbb{E}_t A(k, t) N_t = 2W(t)^2 c_k(t) - 2W(t) \int_{[0,t]} \theta \mathbb{P}_a(Z_0(a) = k) da + W(t) \int_{[0,t]} \theta W(a)^{-1} \mathbb{E}_a [N_a \mathbb{1}_{Z_0(a)=k}] da. \quad (5.1)$$

Moreover, (4.15) entails

$$\psi'(\alpha)^2 \mathbb{E}_t A(k, t) A(l, t) = 2W(t)^2 c_k(t) c_l(t) + RW(t) + o(e^{-\alpha t}),$$

with

$$R := -\psi'(\alpha) \int_0^\infty 2\theta W(a)^{-1} \mathbb{P}_a(Z_0(a) = k) \mathbb{E}_a [A(l, a)] da \\ + \psi'(\alpha) \int_0^\infty 2\theta W(a)^{-1} \mathbb{P}_a(Z_0(a) = l) \mathbb{E}_a [A(k, a)] da \\ + \psi'(\alpha) \int_0^\infty \theta W(a)^{-1} (\mathbb{E}_t [A(k, t) \mathbb{1}_{Z_0(a)=l}] + \mathbb{E}_t [A(l, t) \mathbb{1}_{Z_0(a)=k}]) da.$$

These identities allow us to obtain

$$\begin{aligned}
& \mathbb{E}_t [(A(k, t) - c_k N_t) (A(l, t) - c_l N_t)] = 2W(t)^2 c_k(t) c_l(t) + e^{-\alpha t} R + o(e^{-\alpha t}), \\
& - 2c_l c_k(t) W(t)^2 + 2c_l W(t) \int_{[0, t]} \theta \mathbb{P}_a (Z_0(a) = k) da - c_l W(t) \int_{[0, t]} \theta W(a)^{-1} \mathbb{E}_a [N_a \mathbf{1}_{Z_0(a)=k}] da \\
& - 2c_k c_l(t) W(t)^2 + 2c_l W(t) \int_{[0, t]} \theta \mathbb{P}_a (Z_0(a) = l) da - c_k W(t) \int_{[0, t]} \theta W(a)^{-1} \mathbb{E}_a [N_a \mathbf{1}_{Z_0(a)=l}] da \\
& + c_k c_l W(t)^2 \left( 2 - \frac{1}{W(t)} \right) \\
& = 2W(t)^2 (c_k(t) - c_l) (c_l(t) - c_k) + e^{-\alpha t} \frac{R}{\psi'(\alpha)} + o(e^{-\alpha t}), \\
& + 2c_l W(t) \int_{[0, t]} \theta \mathbb{P}_a (Z_0(a) = k) da - c_l W(t) \int_{[0, t]} \theta W(a)^{-1} \mathbb{E}_a [N_a \mathbf{1}_{Z_0(a)=k}] da \\
& + 2c_l W(t) \int_{[0, t]} \theta \mathbb{P}_a (Z_0(a) = l) da - c_k W(t) \int_{[0, t]} \theta W(a)^{-1} \mathbb{E}_a [N_a \mathbf{1}_{Z_0(a)=l}] da \\
& - c_k c_l W(t).
\end{aligned}$$

Taking the limit as  $t$  goes to infinity leads to

$$\begin{aligned}
M_{k,l} &:= \lim_{t \rightarrow \infty} \psi'(\alpha)^2 e^{-\alpha t} \mathbb{E}_t [(A(k, t) - c_k N_t) (A(l, t) - c_l N_t)] = R \\
&+ 2\psi'(\alpha) c_l \int_{[0, \infty]} \theta \mathbb{P}_a (Z_0(a) = k) da - \psi'(\alpha) c_l \int_{[0, \infty]} \theta W(a)^{-1} \mathbb{E}_a [N_a \mathbf{1}_{Z_0(a)=k}] da \\
&+ 2\psi'(\alpha) c_l \int_{[0, \infty]} \theta \mathbb{P}_a (Z_0(a) = l) da - \psi'(\alpha) c_k \int_{[0, \infty]} \theta W(a)^{-1} \mathbb{E}_a [N_a \mathbf{1}_{Z_0(a)=l}] da \\
&- \psi'(\alpha) c_k c_l.
\end{aligned} \tag{5.2}$$

Finally, since  $\mathbb{P}(N_t > 0) \sim \frac{\alpha}{b}$ ,

$$\mathcal{M}_{i,j} = M_{k_i, k_j}.$$

## 6 Markovian cases

We can get more information on the unknown covariance matrix  $K$  in the case where the life duration distribution is exponential. Our study also cover the case  $\mathbb{P}_V = \delta_\infty$  (Yule case), although it does not fit the conditions required by the Theorem 3.1. The reason comes from our method of calculation for  $\mathbb{E}[A(k, t)\mathcal{E}]$ . Let us consider the filtration  $(\mathcal{F}_t)_{t \in \mathbb{R}_+}$ , where  $\mathcal{F}_t$  is the  $\sigma$ -field generated by the tree truncated above  $t$  and the restriction of the mutation measure on  $[0, t)$ .

Then  $N_t$  is Markovian with respect to  $\mathcal{F}_t$  and for all positive real numbers  $t \leq s$ ,

$$\mathbb{E}[A(k, t) N_s \mid \mathcal{F}_t] = A(k, t) N_t \mathbb{E}[N_{s-t}].$$

So that,

$$\mathbb{E}[A(k, t) N_s] = \mathbb{E}[A(k, t) N_t] (W(s-t) - \mathbb{P}_V \star W(s-t)).$$

By making a renormalization by  $e^{-\alpha s}$  and taking the limit as  $s$  goes to infinity, we get,

$$\mathbb{E}[A(k, t)\mathcal{E}] = \psi'(\alpha)e^{-\alpha t}\mathbb{E}[A(k, t)N_t],$$

since, in the Markovian case, it is known from [8] that

$$\frac{\alpha}{b} = \psi'(\alpha).$$

Suppose first that  $d > 0$ . It follows that,

$$\begin{aligned} \mathbb{E}[(\psi'(\alpha)A(k, t) - e^{\alpha t}c_k\mathcal{E})(\psi'(\alpha)A(l, t) - e^{\alpha t}c_l\mathcal{E})] &= \psi'(\alpha)^2\mathbb{E}_t[A(k, t)A(l, t)]\mathbb{P}(N_t > 0) \\ &\quad - c_k\psi'(\alpha)^2\mathbb{E}_t[A(l, t)N_t]\mathbb{P}(N_t > 0) - c_l\psi'(\alpha)^2\mathbb{E}_t[A(k, t)N_t]\mathbb{P}(N_t > 0) \\ &\quad + 2\psi'(\alpha)e^{2\alpha t}c_kc_l \end{aligned}$$

By (4.4),

$$\mathbb{P}(N_t > 0) = \psi'(\alpha) + \psi'(\alpha)^2\mu e^{-\alpha t} + o(e^{-\alpha t}),$$

so

$$\begin{aligned} &\mathbb{E}[(\psi'(\alpha)A(k, t) - e^{\alpha t}c_k\mathcal{E})(\psi'(\alpha)A(l, t) - e^{\alpha t}c_l\mathcal{E})] \\ &= \mathbb{P}(N_t > 0)\psi'(\alpha)^2\mathbb{E}_t[(A(k, t) - c_kN_t)(A(l, t) - c_lN_t)] + c_kc_l\psi'(\alpha)(2e^{2\alpha t} - \psi'(\alpha)\mathbb{E}_t[N_t^2]\mathbb{P}(N_t > 0)). \end{aligned}$$

Finally, since, using Proposition 2.3,

$$\lim_{t \rightarrow \infty} e^{-\alpha t}(2e^{2\alpha t} - \psi'(\alpha)\mathbb{E}_t[N_t^2]\mathbb{P}(N_t > 0)) = \psi'(\alpha)(1 - 6\mu),$$

it follows from (5.2),

$$\begin{aligned} &\lim_{t \rightarrow \infty} \mathbb{E}[(\psi'(\alpha)A(k, t) - e^{\alpha t}c_k\mathcal{E})(\psi'(\alpha)A(l, t) - e^{\alpha t}c_l\mathcal{E})] \\ &= \psi'(\alpha)M_{k,l} + c_kc_l\psi'(\alpha)^2(1 - 6\mu) = \psi'(\alpha)M_{k,l} + c_kc_l\psi'(\alpha)^2\left(1 - 6\frac{d}{\alpha}\right), \end{aligned}$$

using that  $\mu = \frac{1}{b\mathbb{E}V - 1}$ . In the Yule case, an easy adaptation of the preceding proof leads to

$$\lim_{t \rightarrow \infty} \mathbb{E}[(\psi'(\alpha)A(k, t) - e^{\alpha t}c_k\mathcal{E})(\psi'(\alpha)A(l, t) - e^{\alpha t}c_l\mathcal{E})] = M_{k,l} + c_kc_l.$$

## 7 Numerical studies

The purpose of this section is to analyze our approximation method and the estimation of the error by virtue of numerical experiments. There are several practical difficulties appearing when one tries to perform such study.

The first problem, which involves no conceptual difficulties, lies only on the implementation of the formulas appearing in Theorems 3.1, 3.3 and 3.5. In particular, the computation of the moments of type  $\mathbb{E}[A(k, t)\mathbb{1}_{Z_0(t)=l}]$  are particularly complicated (see Proposition 5.4 in [5]).

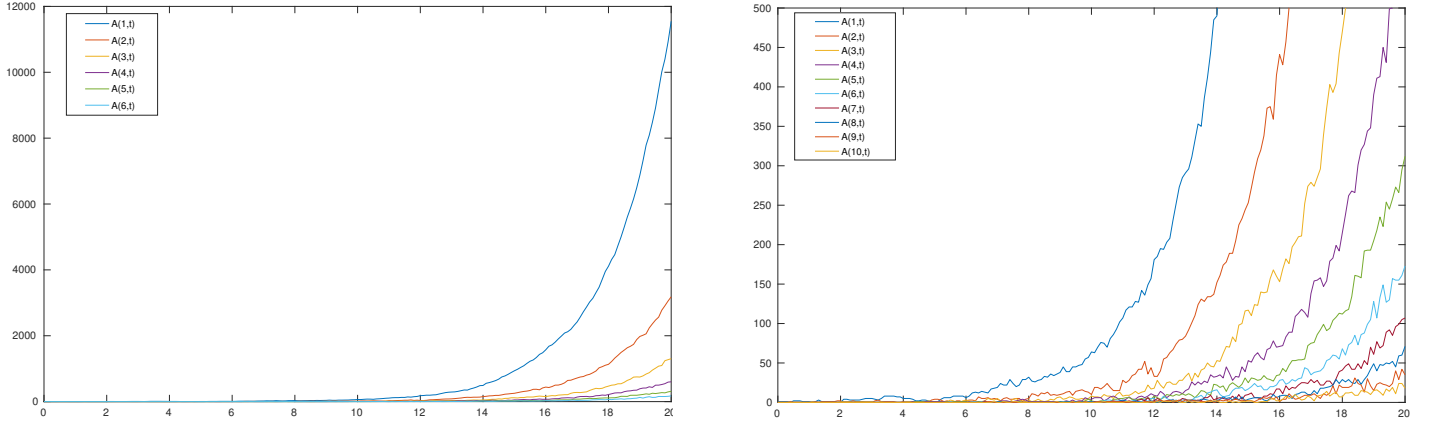


Figure 3: A simulation of the evolution of the frequency spectrum under the given model.

Another difficulty is to obtain numerical approximations of the scale functions  $W$  and  $W_\theta$ . For instance, these functions appear in the computation of the covariance matrix of Theorems 3.3 and 3.5 or when one wants to simulate the *coalescent point process*. To obtain such approximations, we need to apply numerically the Laplace inversion operator to the functions  $\frac{1}{\psi}$  and  $\frac{1}{\psi_\theta}$ .

Unfortunately, the Laplace numerical inversion is a rather difficult problem (see for instance [1] or [2]) which is often computationally expensive. As a consequence, the computational cost of performing multiple numerical integration involving  $W$  or  $W_\theta$  can be important when done with a crude method. Moreover, these methods presents rough numerical instabilities when the original function is exponentially increasing (inverting  $\lambda \rightarrow \frac{1}{1-\lambda}$ , whose inverse is  $x \rightarrow e^x$ , is already a tough numerical problem).

For all these reasons, we provide with this work a Matlab toolbox which handle all these difficulties and allows users to perform numerical experiments without having to take care of these issues.

In this whole section, we are interested in the approximation of the frequency spectrum at a fixed time  $t$  by the sequence  $N_t(c_k)_{k \geq 1}$  (we recall that  $c_k$  was defined in equation (2.13)). As a consequence, the error in this approximation are computed thanks to Theorem 3.5. The parameters of the model are set as follows:

- $\mathbb{P}_V$  is a Rice distribution with shape parameter 1 and scale parameter 1.
- $b = 1$ .
- $\theta = 1$ .

For such parameters  $\alpha$  approximately equals to 0.5. Figure 3 shows the evolution of the frequency spectrum (for  $k$  between 1 and 10) through time. The different quantities seem to growth exponentially with rate  $\alpha$  with a time-shift which depend on  $k$ . An interesting open question would be to understand the behavior of these shifts. In order to stress our methods of approximation, the first



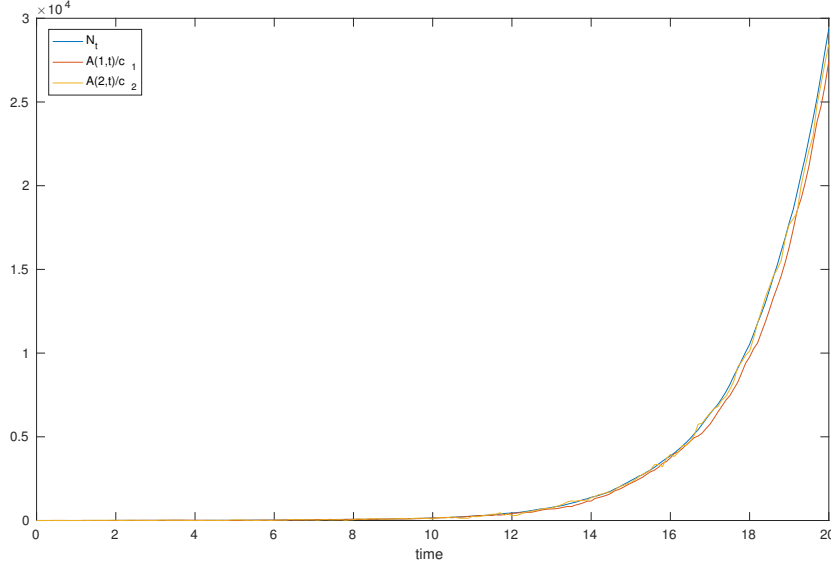


Figure 4: Evolution of the renormalized frequency spectrum  $(A(k, t)/c_k)_{k \geq 1}$  under the given model.

idea is to look to the renormalized frequency spectrum  $\left(\frac{A(k, t)}{c_k}\right)_{k \geq 1}$  which is expected to look like  $(N_t, t \in \mathbb{R}_+)$ . As showed in Figure 4, the approximation seems to be quite accurate for  $k = 1, 2$ . However, a more quantitative analysis is required. Figure 5 shows the absolute error in the approximation of  $A(1, t)$  by  $c_1 N_t$ . This error is a little disappointing since it since to diverge when  $t$  goes to infinity. However, even if, according to Figure 5, the absolute error at time 20 is of order  $10^3$ , the relative error shows that this error is quite small with respect to the value of  $A(1, 20)$ . Another question is about the speed of convergence in the central limit theorem stated in Theorem 3.5. The red curve of Figure 6 shows the density of the Laplace distribution given in Theorem 3.5 in the case of  $A(1, t)$  whereas the blue histogram shows the distribution of  $\psi'(\alpha)(e^{\alpha \frac{t}{2}}(A(1, t) - c_k N_t))$  for  $t = 10$  ( $\alpha t \sim 5$  and  $\mathbb{E}_t[N_t] \sim 300$ ) from 10000 simulations. This Figure highlights the fact that even if the taken time  $t$  is quite small the distribution  $\psi'(\alpha)(e^{\alpha \frac{t}{2}}(A(1, t) - c_k N_t))$  seems already close to the limiting distribution. Figure 7 shows the same kind of behavior in the multidimensional case. To be more quantitative, Figure 8 shows the evolution in time of the distance between the density of limit distribution given in Theorem 3.5 and a kernel estimation of the distribution of  $\psi'(\alpha)(e^{\alpha \frac{t}{2}}(A(1, t) - c_k N_t))$  (the estimation is made from 10000 simulations at each time). This suggest an exponential rate of convergence in Theorem 3.5. In the view of Figure 8, one may think that Berry-Essen type results for Theorem 3.5 would be quite interesting, in particular to understand how the speed of convergence is related to choice of the parameters. Another interesting question which could be partially probed by simulation is the study of the behavior of the error in the clonal supercritical case. Figure 9 shows a kernel estimation (from 10000 simulation) of the density of  $\psi'(\alpha)(e^{\alpha \frac{t}{2}}(A(1, t) - c_k N_t))$  in the clonal supercritical case ( $\theta = 0.2 < \alpha$ ). Figure 9 suggest a totally different behavior with a limit distribution which is asymmetric with respect to 0. In particular, in the view of the shape of the distribution, one could conjecture that the limit is a skew stable distribution.

To end this section, let us goes back to one of the motivation of this work. The following discussion dot not claim to be rigorous and is essentially formal. We recall that the Extended

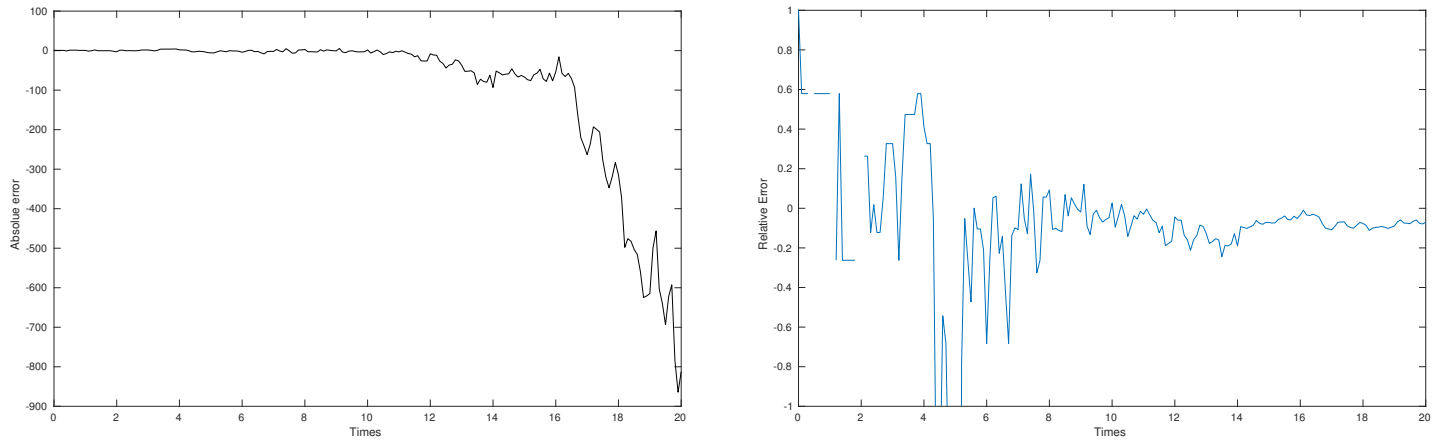


Figure 5: Absolute error (left picture) and relative (right picture) in the approximation of  $A(1, t)$  by  $c_1 N_t$ .

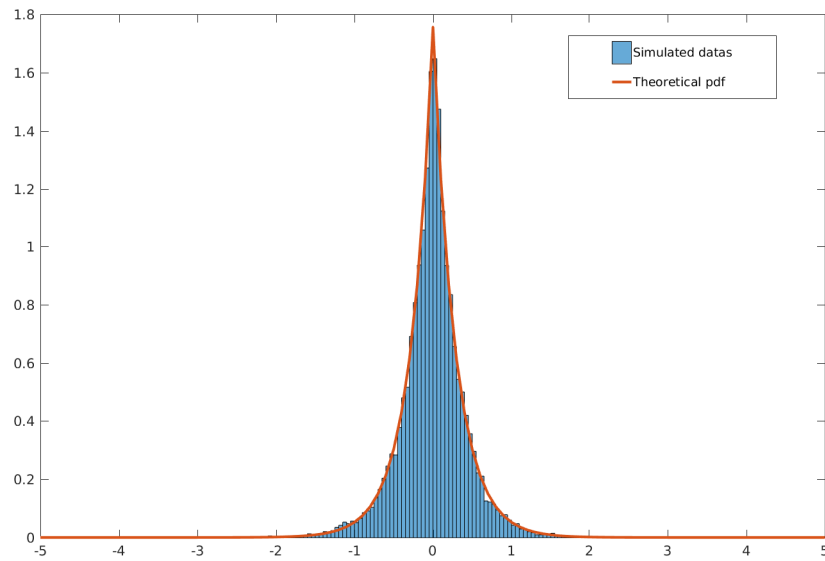


Figure 6: Distribution of the renormalized error and expected limit distribution given by our CLT.

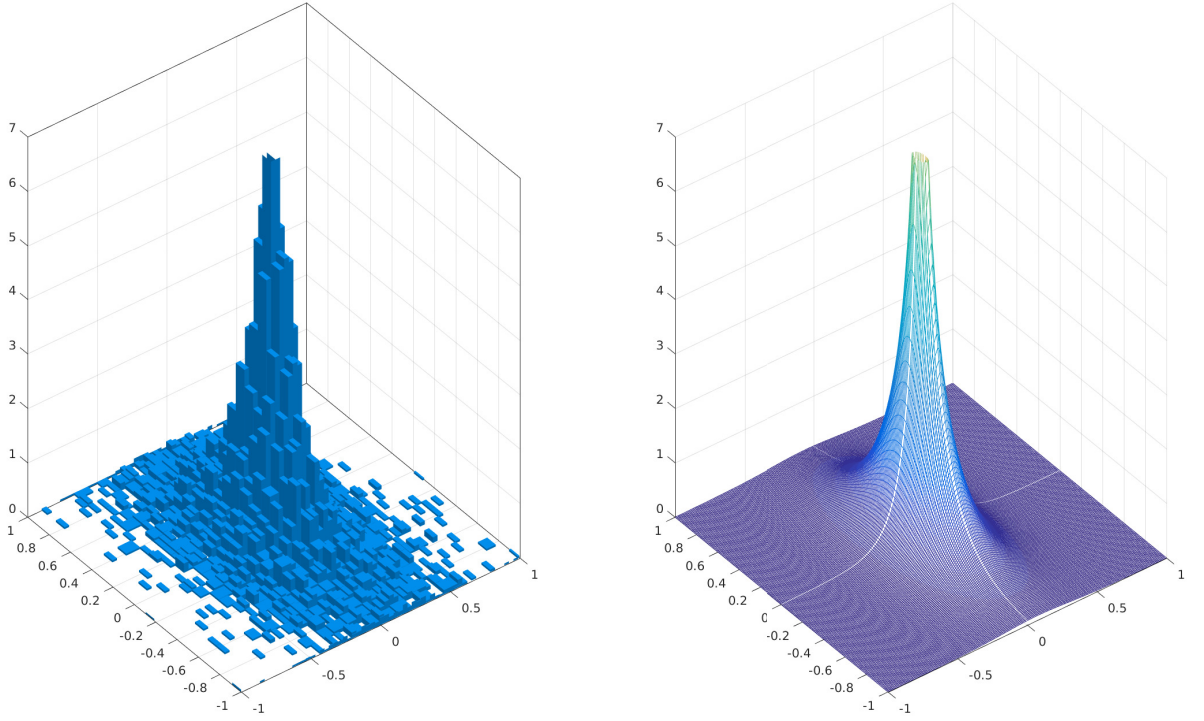


Figure 7: Joint distribution of the renormalized error (left figure) and expected limit distribution (right figure) given by our CLT.

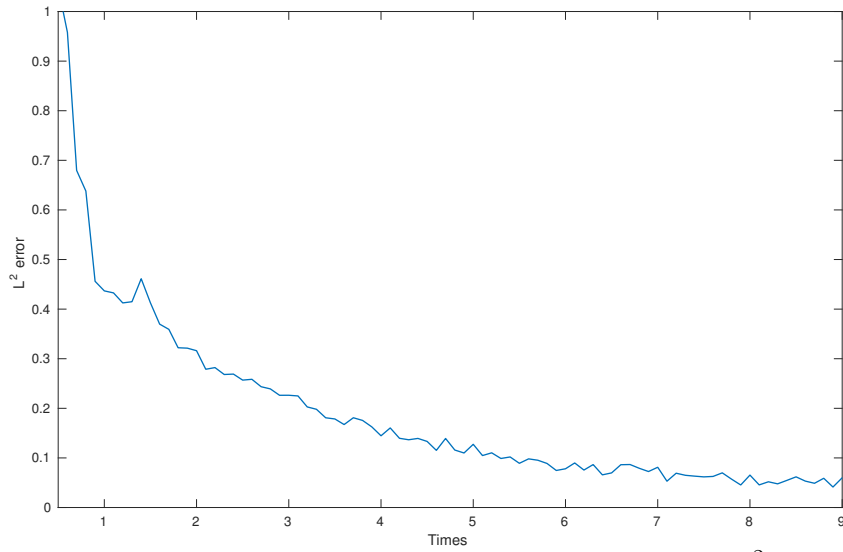


Figure 8: Estimation of the rate of convergence in  $L^2$  norm.

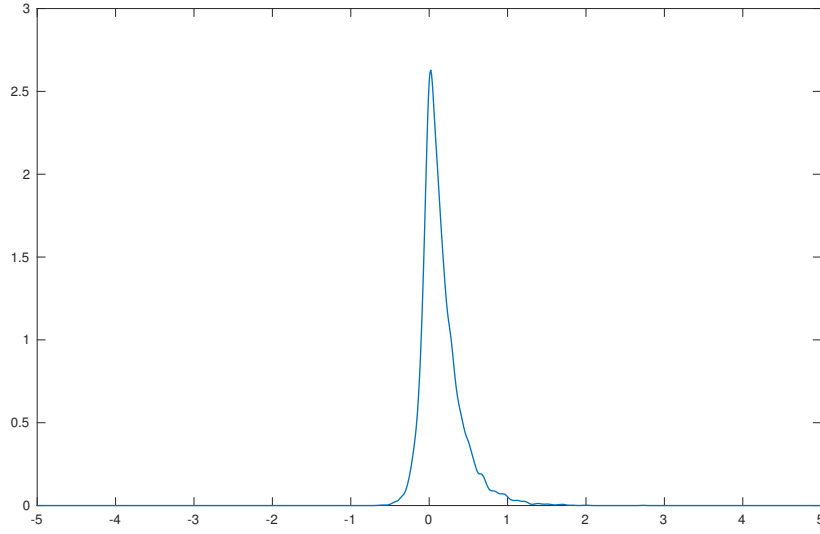


Figure 9: Kernel estimate of the probability density function of the limit distribution in the clonal supercritical case.

Haplotype Homozygosity (EHH) can be used to detect positive selection in a population [18]. In particular, the behavior of the frequency spectrum in this model gives a standard for the behavior of a subpopulation carrying a common allele under neutral evolution. In order to have a rigorous model to describe this phenomenon, we need to introduce a new mutation measure which is different from the one given in Section 2. We define it directly on the CPP but this could be equivalently defined on the splitting tree. So let  $\mathcal{P}$  be a Poisson random measure on  $[0, t] \times \mathbb{N} \times \mathbb{R}_+$  with intensity measure  $\lambda \otimes C \otimes \lambda$ , where  $C$  is the counting measure on  $\mathbb{N}$ , then, for any mutation rate  $\theta$  in  $\mathbb{R}_+$ , we define the  $\theta$ -mutation random measure  $\mathcal{N}_\theta$  by

$$\mathcal{N}_\theta(A \times B) = \int_{A \times B \times [0, \theta]} \mathbf{1}_{H_i > t-a} \mathbf{1}_{i < \mathcal{N}_t} \mathcal{P}(di, da, dx),$$

where, as before, an atom at  $(a, i)$  means that the  $i$ th branch experiences a mutation at time  $t - a$ . This construction allows to increase the mutation in consistent manner. This allows to model the type of an individual at a distance  $x$  (such that the mutation rate is a function of  $x$ ) from the core haplotype (we refer the reader to [18] for more details). Now, following [5], we can define the frequency spectrum at mutation rate  $\theta$  by

$$A^\theta(k, t) = \int_{[0, t] \times \mathbb{N}} \mathbf{1}_{Z_0(i, a)=k} \mathcal{N}_\theta(di, da),$$

where  $Z_0(i, a)$  is the number of individual at time  $t$  carrying the type of the  $i$ th individual at time  $t - a$  (see [5] for more details). Let us also define  $Z_0^\theta(t)$  the number of individuals carrying the type of the first individual at time 0 when the mutation measure is given by  $\mathcal{N}_\theta$ . As expected, the allelic partition of the population becomes thinner as  $\theta$  growth.

Now, the definition of the EHH  $G_\theta(t)$  is the probability that two uniformly sampled individuals in the population have the same type, that is

$$G_t(\theta) = \frac{Z_0^\theta(t)(Z_0^\theta(t) - 1) + \sum_{k \geq 1} k(k-1)A^\theta(k, t)}{N_t(N_t - 1)}.$$

Using that

$$N_t = Z_0^\theta(t) + \sum_{k \geq 1} kA^\theta(k, t),$$

this rewrite

$$G_t(\theta) = \frac{(N_t - \sum_{k \geq 1} kA^\theta(k, t))(N_t - \sum_{k \geq 1} kA^\theta(k, t) - 1) + \sum_{k \geq 1} k(k-1)A^\theta(k, t)}{N_t(N_t - 1)}.$$

Finally, using the approximation

$$(A(k, t))_{k \geq 1} \approx (c_k)_{k \geq 1} N_t$$

proposed in this work, one could expect that

$$G_t(\theta) \approx \frac{\sum_{k \geq 1} k(k-1)c_k}{N_t} = \frac{\int_0^\infty 2\theta e^{-\theta x}(W_\theta(x) - 1)dx}{N_t}.$$

We stress the fact that the above expression make sens only in the clonal subcritical case (in the other cases the integral is not finite). Now, we can look at the accuracy of this approximation in view of numerical simulation. Figure 10 shows the value of the EHH (when  $\theta$  increase) from a simulation of the model (blue curve) and the one obtained using our approximation (red curve). In view of Figure 10, the approximation seems pretty accurate. In order to be more quantitative, Figure 11 shows the relative error between the EHH and its approximation for one simulation. This shows that the error, as least for sufficiently large  $\theta$ , remains under 8%. To end, let us highlight that Theorem 3.5 can be used to give confidence intervals for fixed  $\theta$  but in order to construct tests of selection from curves like these of Figure 10 one would need to have functional CLT in long time for the process  $((A^\theta(k, t) - c_k^\theta N_t)_{k \geq 1}, \theta \in \mathbb{R}_+)$ .

## A A bit of renewal theory

The purpose of this part is to recall some facts on renewal equations borrowed from [10]. Let  $h : \mathbb{R} \rightarrow \mathbb{R}$  be a function bounded on finite intervals with support in  $\mathbb{R}_+$  and  $\Gamma$  a probability measure on  $\mathbb{R}_+$ . The equation

$$F(t) = \int_{\mathbb{R}_+} F(t-s)\Gamma(ds) + h(t),$$

called a renewal equation, is known to admit a unique solution finite on bounded interval.

Here, our interest is focused on the asymptotic behavior of  $F$ . We said that the function  $h$  is DRI (directly Riemann integrable) if for any  $\delta > 0$ , the quantities

$$\delta \sum_{i=0}^n \sup_{t \in [\delta i, \delta(i+1))} f(t)$$

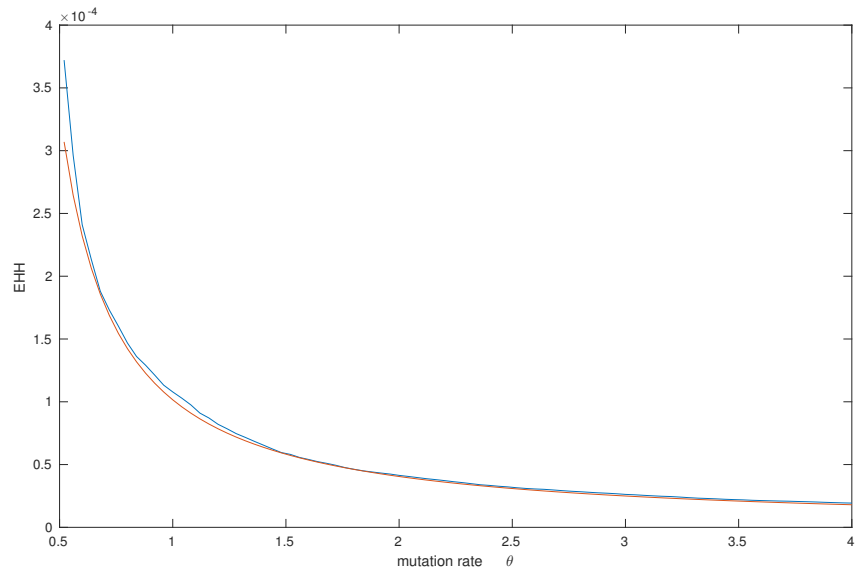


Figure 10: Extended Haplotype Homozygosity (EHH) with the given approximation (red curve) and from simulated data (blue curve)

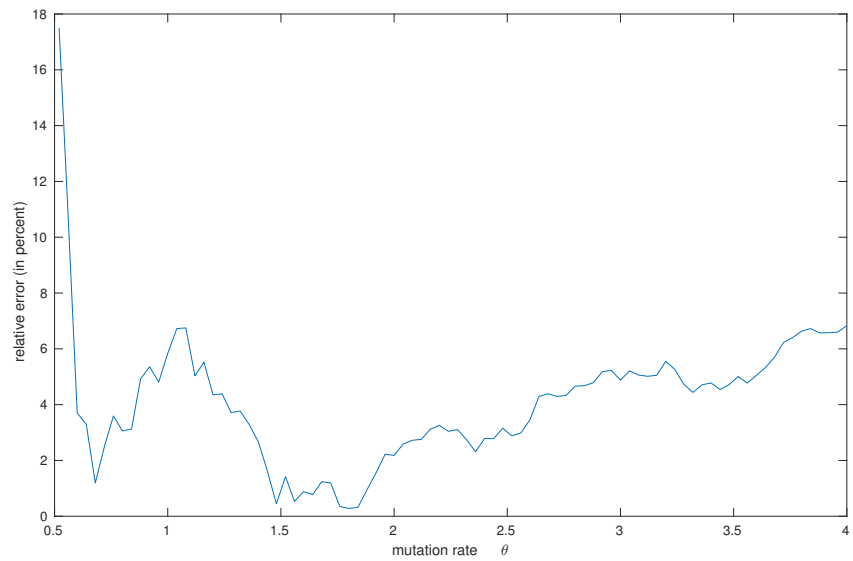


Figure 11: Relative error in the approximation of the EHH

and

$$\delta \sum_{i=0}^n \inf_{t \in [\delta i, \delta(i+1))} f(t)$$

converge as  $n$  goes to infinity respectively to some real number  $I_{sup}^\delta$  and  $I_{inf}^\delta$ , and

$$\lim_{\delta \rightarrow 0} I_{sup}^\delta = \lim_{\delta \rightarrow 0} I_{inf}^\delta < \infty.$$

In the sequel, we use the two following criteria for the DRI property:

**Lemma A.1.** *Let  $h$  a function as defined previously. If  $h$  satisfies one of the next two conditions, then  $h$  is DRI:*

1.  *$h$  is non-negative decreasing and classically Riemann integrable on  $\mathbb{R}_+$ ,*
2.  *$h$  is càdlàg and bounded by a DRI function.*

We can now state the next result, which is constantly used in the sequel.

**Theorem A.2.** *Suppose that  $\Gamma$  is non-lattice, and  $h$  is DRI, then*

$$\lim_{t \rightarrow \infty} F(t) = \gamma \int_{\mathbb{R}_+} h(s) ds,$$

with

$$\gamma := \left( \int_{\mathbb{R}_+} s \Gamma(ds) \right)^{-1},$$

if the above integral is finite, and zero otherwise.

**Remark A.3.** *In particular, if we suppose that  $\Gamma$  is a measure with mass lower than 1, and that there exists a constant  $\alpha \geq 0$  such that*

$$\int_{\mathbb{R}_+} e^{\alpha t} \Gamma(dt) = 1,$$

then, one can perform the change a measure

$$\tilde{\Gamma}(dt) = e^{\alpha t} \Gamma(dt),$$

in order to apply Theorem A.2 to a new renewal equation to obtain the asymptotic behavior of  $F$ . (See [10] for details). This method is also used in the sequel.

## B Formula for the fourth moment of the error

**Lemma B.1.**

$$\begin{aligned}
\mathbb{E}_t \left[ (A(k, t) - c_k N_t)^4 \right] &= 4 \int_{[0, t]} \theta \frac{W(t)}{W(a)} \mathbb{E}_a \left[ \mathbb{1}_{Z_0(a)=k} (A(k, a) - c_k N_a)^3 \right] da \\
&+ 48 \int_{[0, t]} \theta \frac{W(t)^2}{W(a)^2} \left( 1 - \frac{W(a)}{W(t)} \right) \mathbb{E}_a \left[ \mathbb{1}_{Z_0(a)=k} N_a A(k, a) \right] \mathbb{E}_a \left[ (c_k N_a - A(k, a)) \right] da \\
&+ 24 \int_{[0, t]} \theta \frac{W(t)^2}{W(a)^2} \left( 1 - \frac{W(a)}{W(t)} \right) \mathbb{E}_a \left[ \mathbb{1}_{Z_0(a)=k} N_a^2 \right] \mathbb{E}_a \left[ (A(k, a) - c_k N_a) \right] da \\
&+ 24 \int_{[0, t]} \theta \frac{W(t)^2}{W(a)^2} \left( 1 - \frac{W(a)}{W(t)} \right) \mathbb{E}_a \left[ \mathbb{1}_{Z_0(a)=k} A(k, a)^2 \right] \mathbb{E}_a \left[ (A(k, a) - c_k N_a) \right] da \\
&+ 8 \int_{[0, t]} \theta \frac{W(t)^2}{W(a)^2} \left( 1 - \frac{W(a)}{W(t)} \right) \mathbb{P}_a (Z_0(a) = k) \mathbb{E}_a \left[ (A(k, a) - c_k N_a)^3 \right] da \\
&+ 48 \int_{[0, t]} \theta \frac{W(t)^2}{W(a)^2} \left( 1 - \frac{W(a)}{W(t)} \right) \mathbb{E}_a \left[ \mathbb{1}_{Z_0(a)=k} A(k, a) \right] \mathbb{E}_a \left[ (A(k, a) - c_k N_a)^2 \right] da \\
&+ 72 \int_{[0, t]} \theta \frac{W(t)^3}{W(a)^3} \left( 1 - \frac{W(a)}{W(t)} \right)^2 \mathbb{E}_a \left[ \mathbb{1}_{Z_0(a)=k} (A(k, a) - c_k N_a) \right] \mathbb{E}_a \left[ (A(k, a) - c_k N_a)^2 \right] da \\
&+ 72 \int_{[0, t]} \theta \frac{W(t)^3}{W(a)^3} \left( 1 - \frac{W(a)}{W(t)} \right)^2 \mathbb{P}_a (Z_0(a) = k) \mathbb{E}_a \left[ (A(k, a) - c_k N_a)^2 \right] \mathbb{E}_a \left[ A(k, a) - N_a c_k \right] da \\
&+ 96 \int_{[0, t]} \theta \frac{W(t)^4}{W(a)^4} \left( 1 - \frac{W(a)}{W(t)} \right)^3 \mathbb{P}_a (Z_0(a) = k) \mathbb{E}_a \left[ (A(k, a) - c_k N_a)^3 \right] da + c_k^4 \mathbb{E}_t N_t^4
\end{aligned}$$

*Proof.* The proof of this Lemma lies on the calculation of the expectation of each term in the development of

$$(A(k, t) - c_k N_t)^4.$$

To make this, we intensively use the relation (2.10) and the method developed in [5]. We begin by computing

$$\mathbb{E}_t \left[ A(k, t)^4 \right].$$



Formula (2.10) gives us,

$$\begin{aligned}
A(k, t)^4 &= 4 \int_{[0, t] \times \mathbb{N}} \mathbb{1}_{Z_0^i(a)=k_i} \sum_{u_{1:3}=1}^{N_{t-a}^{(t)}} \prod_{\substack{j=1 \\ i \neq j}}^3 A^{(u_j)}(k, a) \mathcal{N}(da, di) \\
&= 4 \int_{[0, t] \times \mathbb{N}} \mathbb{1}_{Z_0^i(a)=k} A^i(k, a) A^i(k, a) A^i(k, a) \mathcal{N}(da, di) \\
&\quad + 4 \int_{[0, t] \times \mathbb{N}} \mathbb{1}_{Z_0^i(a)=k} \sum_{\substack{j_1, j_2, j_3=1 \\ j_1 \neq j_2 \neq j_3 \neq i}}^{N_{t-a}^{(t)}} A^{j_1}(k, a) A^{j_2}(k, a) A^{j_3}(k, a) \mathcal{N}(da, di) \\
&\quad + 12 \int_{[0, t] \times \mathbb{N}} \mathbb{1}_{Z_0^i(a)=k} A^i(k, a) A^i(k, a) \sum_{j=1, j \neq i}^{N_{t-a}^{(t)}} A^j(k, a) \mathcal{N}(da, di) \\
&\quad + 4 \int_{[0, t] \times \mathbb{N}} \mathbb{1}_{Z_0^i(a)=k} \sum_{j=1, j \neq i}^{N_{t-a}^{(t)}} A^j(k, a)^3 \mathcal{N}(da, di) \\
&\quad + 12 \int_{[0, t] \times \mathbb{N}} \mathbb{1}_{Z_0^i(a)=k} A^i(k, a) \sum_{j_1, j_2=1, j_1 \neq j_2 \neq i}^{N_{t-a}^{(t)}} A^{j_1}(k, a) A^{j_2}(k, a) \mathcal{N}(da, di) \\
&\quad + 24 \int_{[0, t] \times \mathbb{N}} \mathbb{1}_{Z_0^i(a)=k} A^i(k, a) \sum_{j_1=1, j_1 \neq i}^{N_{t-a}^{(t)}} A^{j_1}(k, a) A^{j_1}(k, a) \mathcal{N}(da, di) \\
&\quad + 12 \int_{[0, t] \times \mathbb{N}} \mathbb{1}_{Z_0^i(a)=k} \sum_{j_1, j_2=1, j_1 \neq j_2 \neq i}^{N_{t-a}^{(t)}} A^{j_1}(k, a)^2 A^{j_2}(k, a) \mathcal{N}(da, di). \tag{B.1}
\end{aligned}$$

The decomposition of the sum in form

$$\sum_{u_{1:3}=1}^{N_{t-a}^{(t)}},$$

has then been made to distinguish independence properties in our calculation. Actually, as soon as,  $i \neq j$ ,  $A^i(k, a)$  is independent from  $A^j(k, a)$  (see [5] for details). It is essential to note that the expectation of these integrals with respect to the random measure  $\mathcal{N}$  are all calculated thanks to

Theorem 3.1 of [5]. So, taking the expectation now leads to,

$$\begin{aligned}
\mathbb{E}_t [A(k, t)^4] = & 4 \int_{[0, t]} \theta \mathbb{E}_a [N_{t-a}^{(t)}] \mathbb{E}_a [\mathbb{1}_{Z_0(a)=k} A(k, a)^3] \theta da \\
& + 4 \int_{[0, t]} \theta \mathbb{P}_a (Z_0(a) = k) \mathbb{E}_a \left[ \binom{N_{t-a}^{(t)}}{(4)} \right] \mathbb{E}_a [A(k, a)]^3 da \\
& + 12 \int_{[0, t]} \theta \mathbb{E}_a [\mathbb{1}_{Z_0(a)=k} A(k, a)^2] \mathbb{E}_a \left[ \binom{N_{t-a}^{(t)}}{(2)} \right] \mathbb{E}_a [A(k, a)] da \\
& + 4 \int_{[0, t]} \theta \mathbb{P}_a (Z_0(a) = k) \mathbb{E}_a \left[ \binom{N_{t-a}^{(t)}}{(2)} \right] \mathbb{E}_a [A(k, a)]^3 da \\
& + 12 \int_{[0, t]} \theta \mathbb{E}_a [\mathbb{1}_{Z_0(a)=k} A(k, a)] \mathbb{E}_a \left[ \binom{N_{t-a}^{(t)}}{(3)} \right] \mathbb{E}_a [A(k, a)]^2 da \\
& + 24 \int_{[0, t]} \theta \mathbb{E}_a [\mathbb{1}_{Z_0(a)=k} A(k, a)] \mathbb{E}_a \left[ \binom{N_{t-a}^{(t)}}{(2)} \right] \mathbb{E}_a [A(k, a)^2] da \\
& + 12 \int_{[0, t]} \theta \mathbb{P}_a (Z_0(a) = k) \mathbb{E}_a \left[ \binom{N_{t-a}^{(t)}}{(3)} \right] \mathbb{E}_a [A(k, a)^2] \mathbb{E}_a [A(k, a)] da.
\end{aligned}$$

Using the same method for all the other terms and that, for any positive real number  $a$  lower than  $t$ ,

$$N_t = \sum_{i=1}^{N_{t-a}^{(t)}} N_a^{(i)},$$

we get Lemma B.1 by reassembling similar terms together. The last term is obtained using the geometric distribution of  $N_t$  under  $\mathbb{P}_t$ .  $\square$

## C Boundedness of the fourth moment

**Lemma C.1.** *We begin the proof of the boundedness of the fourth moment by some estimates.*

$$\mathbb{E}_t [(A(k, t) - c_k N_t)] = \mathcal{O} \left( e^{-(\theta-\alpha)t} \right), \quad (\text{i})$$

$$\mathbb{E}_t \left[ (A(k, t) - c_k N_t)^3 \right] = \mathcal{O} (W(t)^2), \quad (\text{ii})$$

$$\mathbb{E}_t \left[ (A(k, t) - c_k N_t)^2 \right] = \mathcal{O} (W(t)), \quad (\text{iii})$$

$$\mathbb{E}_t N_t^n = \mathcal{O}(e^{n\alpha t}), \quad n \in \mathbb{N}^*, \quad (\text{iv})$$

$$\mathbb{P}_t (Z_0(t) = k) = \mathcal{O}(e^{(\alpha-\theta)t}). \quad (\text{v})$$

*Proof.* Relation (i) is easily obtained using the expectation of  $N_t$  and  $A(k, t)$  using (2.11), (2.13) and the behaviour of  $W$  provided by Proposition 2.3. The relation (iii) has been obtained in the proof

of Theorem 6.1 in [5]. The two last relations are easily obtained from (2.4), (2.8) and Lemma 2.2. The relation (ii) is obtained using the following estimation,

$$\left| \mathbb{E}_t \left[ (A(k, a) - c_k N_a)^3 \right] \right| \leq \mathbb{E}_t \left[ N_a (A(k, a) - c_k N_a)^2 \right].$$

We begin the proof by computing the r.h.s. of the previous inequality using the same techniques as in Appendix A.

$$\begin{aligned} \mathbb{E} [A(k, t)^2 N_t] &= 2 \int_0^t \theta \frac{W(t)}{W(a)} \mathbb{E} [N_a A(k, a) \mathbf{1}_{Z_0(a)=k}] da \\ &+ 4 \int_0^t \theta \frac{W(t)^2}{W(a)^2} \left( 1 - \frac{W(a)}{W(t)} \right) \mathbb{E} [N_a \mathbf{1}_{Z_0(a)=k}] \mathbb{E} [A(k, a)] da \\ &+ 4 \int_0^t \theta \frac{W(t)^2}{W(a)^2} \left( 1 - \frac{W(a)}{W(t)} \right) \mathbb{E} [A(k, a) \mathbf{1}_{Z_0(a)=k}] \mathbb{E} [N_a] da \\ &+ 4 \int_0^t \theta \frac{W(t)^2}{W(a)^2} \left( 1 - \frac{W(a)}{W(t)} \right) \mathbb{P}_a (Z_0(a) = k) \mathbb{E} [A(k, a) N_a] da \\ &+ 12 \int_0^t \theta \frac{W(t)^3}{W(a)^3} \left( 1 - \frac{W(a)}{W(t)} \right)^2 \mathbb{P}_a (Z_0(a) = k) \mathbb{E} [A(k, a)] \mathbb{E} [N_a] da. \end{aligned}$$

$$\begin{aligned} 2\mathbb{E} [A(k, t) N_t^2] &= 2 \int_0^t \theta \frac{W(t)}{W(a)} \mathbb{E} [N_a^2 \mathbf{1}_{Z_0(a)=k}] da \\ &+ 8 \int_0^t \theta \frac{W(t)^2}{W(a)^2} \left( 1 - \frac{W(a)}{W(t)} \right) \mathbb{E} [N_a \mathbf{1}_{Z_0(a)=k}] \mathbb{E} [N_a] da \\ &+ 4 \int_0^t \theta \frac{W(t)^2}{W(a)^2} \left( 1 - \frac{W(a)}{W(t)} \right) \mathbb{P}_a (Z_0(a) = k) \mathbb{E} [N_a^2] da \\ &+ 12 \int_0^t \theta \frac{W(t)^3}{W(a)^3} \left( 1 - \frac{W(a)}{W(t)} \right)^2 \mathbb{P}_a (Z_0(a) = k) \mathbb{E} [N_a]^2 da. \end{aligned}$$

Finally,

$$\begin{aligned} \mathbb{E} [N_t (A(k, t) - c_k N_t)^2] &= 2 \int_0^t \theta \frac{W(t)}{W(a)} \mathbb{E} [N_a (A(k, a) - c_k N_a) \mathbf{1}_{Z_0(a)=k}] da \\ &+ 4 \int_0^t \theta \frac{W(t)^2}{W(a)^2} \left( 1 - \frac{W(a)}{W(t)} \right) \mathbb{E} [N_a \mathbf{1}_{Z_0(a)=k}] \mathbb{E} [A(k, a) - c_k N_a] da \\ &+ 4 \int_0^t \theta \frac{W(t)^2}{W(a)^2} \left( 1 - \frac{W(a)}{W(t)} \right) \mathbb{E} [(A(k, a) - c_k N_a) \mathbf{1}_{Z_0(a)=k}] \mathbb{E} [N_a] da \\ &+ 4 \int_0^t \theta \frac{W(t)^2}{W(a)^2} \left( 1 - \frac{W(a)}{W(t)} \right) \mathbb{P}_a (Z_0(a) = k) \mathbb{E} [N_a (A(k, a) - c_k N_a)] da \\ &+ 12 \int_0^t \theta \frac{W(t)^3}{W(a)^3} \left( 1 - \frac{W(a)}{W(t)} \right)^2 \mathbb{P}_a (Z_0(a) = k) \mathbb{E} [N_a] \mathbb{E} [A(k, a) - c_k N_a] da \\ &+ c_k^2 \mathbb{E}_t N_t^3. \end{aligned}$$

Now, an analysis similar to the one of Lemma 4.7 leads to the result.  $\square$

*Proof of Lemma 4.7.* The ideas of the proof, is to analyses one to one every terms of the expression of

$$\mathbb{E}_t \left[ (A(k, t) - c_k N_t)^4 \right],$$

given by Lemma B.1 using Lemma C.1 to show that they behave as  $\mathcal{O}(W(t)^2)$ . Since the ideas are the same for every terms, we just give a few examples.

First of all, we consider

$$\int_{[0,t]} \frac{W(t)}{W(a)} \mathbb{E}_a \left[ \mathbb{1}_{Z_0(a)=k} (A(k, a) - c_k N_a)^3 \right] da.$$

Using Lemma C.1 (ii), we have

$$\int_{[0,t]} \frac{W(t)}{W(a)} \mathbb{E}_a \left[ \mathbb{1}_{Z_0(a)=k} (A(k, a) - c_k N_a)^3 \right] da = \mathcal{O}(W(t)^2).$$

Now take the term

$$\int_{[0,t]} \frac{W(t)^2}{W(a)^2} \mathbb{E}_a \left[ \mathbb{1}_{Z_0(a)=k} N_a^2 \right] \mathbb{E}_a [(A(k, a) - c_k N_a)] da,$$

we have from Lemma C.1 (i) and (iv),

$$\int_{[0,t]} \frac{W(t)^2}{W(a)^2} \mathbb{E}_a \left[ \mathbb{1}_{Z_0(a)=k} N_a^2 \right] \mathbb{E}_a [(A(k, a) - c_k N_a)] da \leq \int_{[0,t]} \frac{W(t)^2}{W(a)^2} \mathbb{E}_a [N_a^2] e^{-(\theta-\alpha)a} da = \mathcal{O}(W(t)^2).$$

Every term in  $W(t)$  or  $W(t)^2$  are treated this way. Now, we consider the term in  $W(t)^4$  which is

$$I := 96 \int_{[0,t]} \frac{W(t)^4}{W(a)^4} \mathbb{P}_a(Z_0(a) = k) \mathbb{E}_a [(A(k, a) - c_k N_a)]^3 da + 24W(t)^4 c_k^4,$$

since  $N_t$  is geometrically distributed under  $\mathbb{P}_t$ , and that

$$\mathbb{E}_t N_t^4 = 24W(t)^4 - 36W(t)^3 + \mathcal{O}(W(t)^2). \quad (\text{C.1})$$

On the other hand, using the law of  $Z_0(t)$  given by (2.8) and the expectation of  $A(k, t)$  given by (2.11) (under  $\mathbb{P}_t$ ), we have,

$$\begin{aligned} & 96 \int_{[0,t]} \frac{W(t)^4}{W(a)^4} \mathbb{P}_a(Z_0(a) = k) \mathbb{E}_a [(A(k, a) - c_k N_a)]^3 da \\ &= -96W(t)^4 \int_0^t \frac{\theta e^{-\theta a}}{W_\theta(a)^2} \left(1 - \frac{1}{W_\theta(a)}\right)^{k-1} \left( \int_0^a \frac{\theta e^{-\theta s}}{W_\theta(s)^2} \left(1 - \frac{1}{W_\theta(s)}\right)^{k-1} ds \right)^3 da \\ &= -24W(t)^4 \left( \int_0^t \frac{\theta e^{-\theta a}}{W_\theta(a)^2} \left(1 - \frac{1}{W_\theta(a)}\right)^{k-1} da \right)^4. \end{aligned}$$

Finally,

$$I = 24W(t)^4 \left( \int_t^\infty \frac{\theta e^{-\theta a}}{W_\theta(a)^2} \left( 1 - \frac{1}{W_\theta(a)} \right)^{k-1} da \right)^4 = \mathcal{O} \left( W(t)^4 e^{-4\theta t} \right) = o(1).$$

The last example is the most technical and relies with the term in  $W(t)^3$ , which is, using (C.1) and Lemma B.1,

$$\begin{aligned} J := & 72 \int_{[0,t]} \frac{W(t)^3}{W(a)^3} \mathbb{E}_a [\mathbb{1}_{Z_0(a)=k} (A(k, a) - c_k N_a)] \mathbb{E}_a [(A(k, a) - c_k N_a)]^2 da \\ & + 72 \int_{[0,t]} \frac{W(t)^3}{W(a)^3} \mathbb{P}_a (Z_0(a) = k) \mathbb{E}_a [(A(k, a) - c_k N_a)^2] \mathbb{E}_a [A(k, a) - N_a c_k] da \\ & - 288 \int_{[0,t]} \frac{W(t)^3}{W(a)^3} \mathbb{P}_a (Z_0(a) = k) \mathbb{E}_a [(A(k, a) - c_k N_a)]^3 da - 36c_k^4 W(t)^3. \end{aligned}$$

On the other hand, using the calculus made in the proof of Theorem 6.3 of [5], we have

$$\begin{aligned} & \mathbb{E}_a [(A(k, a) - c_k N_a)^2] \\ & = 4 \int_{[0,a]} \frac{W(a)^2}{W(s)^2} \left( 1 - \frac{W(s)}{W(a)} \right) \mathbb{P}_s (Z_0(s) = k) \mathbb{E}_a (A(k, s) - c_k N_s) ds \\ & \quad + 2 \int_{[0,a]} \frac{W(s)}{W(a)} \mathbb{E}_a [\mathbb{1}_{Z_0(s)=k} (A(k, s) - c_k N_s)] ds + c_k^2 W(a)^2 \left( 2 - \frac{1}{W(a)} \right). \end{aligned}$$

Substituting this last expression in  $J$  leads to

$$\begin{aligned} J = & -144 \int_{[0,t]} \frac{W(t)^3}{W(a)^3} \mathbb{E}_a [\mathbb{1}_{Z_0(a)=k} (A(k, a) - c_k N_a)] \int_{[a,\infty]} \frac{\mathbb{P}(Z_0(a) = k)}{W(s)^2} \mathbb{E}_a [(A(k, s) - c_k N_s)] ds da \\ & + 144W(t)^3 \int_{[0,t]} \frac{1}{W(a)} \mathbb{E}_a [\mathbb{1}_{Z_0(a)=k} (A(k, a) - c_k N_a)] \int_{[a,t]} \frac{1}{W(s)^2} \mathbb{P}_s (Z_0(s) = k) \mathbb{E}_a [A(k, s) - N_s c_k] da \\ & - 144c_k^2 \int_{[0,t]} \frac{W(t)^3}{W(a)} \mathbb{P}_a (Z_0(a) = k) \mathbb{E}_a [A(k, a) - N_a c_k] da \\ & + 144 \int_{[0,t]} \frac{W(t)^3}{W(a)^3} \mathbb{P} (Z_0(a) = k) \mathbb{E}_a [A(k, a) - N_a c_k]^3 da \\ & - 288 \int_{[0,t]} \frac{W(t)^3}{W(a)^2} \mathbb{P}_a (Z_0(a) = k) \int_{[0,a]} \frac{1}{W(s)} \mathbb{P}_s (Z_0(s) = k) \mathbb{E}_a (A(k, s) - c_k N_s) ds \mathbb{E}_a [A(k, a) - N_a c_k] da \\ & + 72 \int_{[0,t]} \frac{W(t)^3}{W(a)} \mathbb{P}_a (Z_0(a) = k) c_k^2 \left( 2 - \frac{1}{W(a)} \right) \mathbb{E}_a [A(k, a) - N_a c_k] da \\ & - 288 \int_{[0,t]} \frac{W(t)^3}{W(a)^3} \mathbb{P}_a (Z_0(a) = k) \mathbb{E}_a [(A(k, a) - c_k N_a)]^3 da - 36c_k^4 W(t)^3. \end{aligned}$$

Using many times that,

$$\begin{aligned}
& \int_{[0,t]} \frac{\theta \mathbb{P}(Z_0(a) = k)}{W(s)^2} \mathbb{E}_a [(A(k, s) - c_k N_s)] ds \\
&= - \int_{[0,t]} \frac{\theta e^{-\theta s}}{W_\theta(s)^2} \left(1 - \frac{1}{W_\theta(s)}\right)^{k-1} \int_{[s,\infty]} \frac{\theta e^{-\theta u}}{W_\theta(u)^2} \left(1 - \frac{1}{W_\theta(u)}\right)^{k-1} du ds \\
&= \frac{c_k^2}{2} - \frac{1}{2} \left( \int_{[t,\infty]} \frac{\theta e^{-\theta s}}{W_\theta(s)^2} \left(1 - \frac{1}{W_\theta(s)}\right)^{k-1} ds \right)^2,
\end{aligned}$$

thanks to (2.8), (2.11), and (2.6), we finally get

$$\begin{aligned}
J &= -144 (c_k^2 - c_k(t)^2) \int_{[0,t]} \frac{W(t)^3}{W(a)^3} \mathbb{E}_a [\mathbb{1}_{Z_0(a)=k} (A(k, a) - c_k N_a)] da \\
&\quad + 36W(t)^3 \left( c_k^2 \left( \int_{[t,\infty]} \frac{W(t)^3}{W(a)^3} \mathbb{E}_a [A(k, a) - N_a c_k]^3 da \right)^2 - \left( \int_{[t,\infty]} \frac{W(t)^3}{W(a)^3} \mathbb{E}_a [A(k, a) - N_a c_k]^3 da \right)^4 \right) \\
&\quad + 144 (c_k - c_k(t))^2 \int_{[0,t]} \frac{W(t)^3}{W(a)} \mathbb{E}_a [A(k, a) - N_a c_k] da \\
&\quad + 36W(t)^3 (c_k - c_k(t))^4.
\end{aligned}$$

This shows that  $J$  is  $\mathcal{O}(W(t)^2)$ . □

## References

- [1] Joseph Abate, Gagan L Choudhury, and Ward Whitt. An introduction to numerical transform inversion and its application to probability models. In *Computational probability*, pages 257–323. Springer, 2000.
- [2] Joseph Abate and Ward Whitt. A unified framework for numerically inverting laplace transforms. *INFORMS Journal on Computing*, 18(4):408–421, 2006.
- [3] K. B. Athreya and P. E. Ney. *Branching processes*. Dover Publications, Inc., Mineola, NY, 2004. Reprint of the 1972 original [Springer, New York; MR0373040].
- [4] Jean Bertoin. The structure of the allelic partition of the total population for Galton-Watson processes with neutral mutations. *Ann. Probab.*, 37(4):1502–1523, 2009.
- [5] Nicolas Champagnat and Henry Benoit. Moments of the frequency spectrum of a splitting tree with neutral poissonian mutations. *Electron. J. Probab.*, 21:34 pp., 2016.
- [6] Nicolas Champagnat and Amaury Lambert. Splitting trees with neutral Poissonian mutations I: Small families. *Stochastic Process. Appl.*, 122(3):1003–1033, 2012.

- [7] Nicolas Champagnat and Amaury Lambert. Splitting trees with neutral Poissonian mutations II: Largest and oldest families. *Stochastic Process. Appl.*, 123(4):1368–1414, 2013.
- [8] Nicolas Champagnat, Amaury Lambert, and Mathieu Richard. Birth and death processes with neutral mutations. *Int. J. Stoch. Anal.*, pages Art. ID 569081, 20, 2012.
- [9] Warren J. Ewens. *Mathematical population genetics. I*, volume 27 of *Interdisciplinary Applied Mathematics*. Springer-Verlag, New York, second edition, 2004. Theoretical introduction.
- [10] William Feller. *An introduction to probability theory and its applications. Vol. II*. Second edition. John Wiley & Sons, Inc., New York-London-Sydney, 1971.
- [11] J. Geiger and G. Kersting. Depth-first search of random trees, and Poisson point processes. In *Classical and modern branching processes (Minneapolis, MN, 1994)*, volume 84 of *IMA Vol. Math. Appl.*, pages 111–126. Springer, New York, 1997.
- [12] R. C. Griffiths and Anthony G. Pakes. An infinite-alleles version of the simple branching process. *Adv. in Appl. Probab.*, 20(3):489–524, 1988.
- [13] B. Henry. Central limit theorem for supercritical binary homogeneous Crump-Mode-Jagers processes. *to appear in ESAIM:Probability and Statistics*, November 2016.
- [14] Amaury Lambert. The contour of splitting trees is a Lévy process. *Ann. Probab.*, 38(1):348–395, 2010.
- [15] Amaury Lambert, Lea Popovic, et al. The coalescent point process of branching trees. *The Annals of Applied Probability*, 23(1):99–144, 2013.
- [16] Amaury Lambert and Pieter Trapman. Splitting trees stopped when the first clock rings and vervaat’s transformation. *Journal of Applied Probability*, 50(01):208–227, 2013.
- [17] Mathieu Richard. *Arbres, Processus de branchement non Markoviens et processus de Lévy*. Thèse de doctorat, Université Pierre et Marie Curie, Paris 6.
- [18] Pardis C Sabeti, David E Reich, John M Higgins, Haninah ZP Levine, Daniel J Richter, Stephen F Schaffner, Stacey B Gabriel, Jill V Platko, Nick J Patterson, Gavin J McDonald, et al. Detecting recent positive selection in the human genome from haplotype structure. *Nature*, 419(6909):832–837, 2002.